

C U R S O S E C O N G R E S O S

La Filosofía Analítica en el Cambio de Milenio

ACTAS

Santiago de Compostela,
1-4 de diciembre, 1999



EDICIÓN A CARGO DE
José L. Falguera
Uxía Rivas
José M. Sagüillo

UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

publicacións

Proceedings of the International Congress

**Analytic Philosophy
at the Turn of the Millennium**

Actas del Congreso Internacional

**La Filosofía Analítica
en el Cambio de Milenio**

CURSOS E CONGRESOS DA
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Nº 120



Proceedings of the International Congress

***Analytic Philosophy
at the Turn of the Millennium***

Actas del Congreso Internacional

***La Filosofía Analítica
en el Cambio de Milenio***

Santiago de Compostela, 1-4, December / Diciembre, 1999

Edited by / Edición a cargo de
JOSÉ L. FALGUERA • UXÍA RIVAS • JOSÉ M. SAGÜILLO

1999

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CONGRESO INTERNACIONAL LA FILOSOFIA ANALITICA EN EL CAMBIO DE MILENIO (1999. Santiago de Compostela)

Proceedings of the International Congress Analytic Philosophy at the Turn of the Millennium: 1-4 December, 1999, Santiago de Compostela (Spain) / edited by José L. Falguera, Uxía Rivas, José M. Sagüillo = Actas del Congreso Internacional La Filosofía Analítica en el Cambio de Milenio : 1-4 de Diciembre de 1999, Santiago de Compostela (España) / Edición a cargo de José L. Falguera, Uxía Rivas, José M. Sagüillo. -- Santiago de Compostela : Universidade, Servicio de Publicacións e Intercambio Científico, 1999. -- 708 p. ; 24 cm. -- (Cursos e Congresos da Universidade de Santiago de Compostela ; 120). -- Bibliografía. -- D.L.C. 2181/99. -- ISBN 84-8121-786-7

I.Filosofía analítica-Congresos. I.Falguera, José L.,ed.lit. II.Rivas, Uxía, ed.lit. III.Sagüillo, José M., ed.lit. IV.Universidade de Santiago de Compostela. Servicio de Publicacións e Intercambio Científico, ed. V.Título. VI.Serie
167/168 (063)

© Universidade de Santiago de Compostela, 1999

Deseño de cuberta

Alejandro Vidal

Maquetación

Santiago Fernández Lanza

Edita

Servicio de Publicacións
da Universidade de Santiago de Compostela
Campus universitario sur

Imprime

Imprenta Universitaria
Campus universitario sur

ISBN 84-8121-786-7

Dep. Legal: C-2181-1999

International Congress / Congreso Internacional

Analytic Philosophy at the Turn of the Millennium
La Filosofía Analítica en el Cambio de Milenio

1-4 December 1999 / 1-4 de Diciembre de 1999

Santiago de Compostela (Spain/España)

Organised by / Organiza:

Sociedad Española de Filosofía Analítica (S.E.F.A.)
Área de Lógica y Filosofía de la Ciencia de la U.S.C.

Supported by / Apoya:

European Society for Analytic Philosophy (E.S.A.P.)

Scientific committee/ Comité científico:

J.J. Acero (Univ. de Granada); J. Blasco (Univ. de Valencia); J. Corbi (Univ. de Valencia); J. Ezquerro (Univ. del País Vasco); L. Fernández-Moreno (Univ. Complutense de Madrid); M. García Carpintero (Univ. de Barcelona); R. Giere (Univ. of Minnesota); M. Liz (Univ. de La Laguna); C. Martínez-Vidal (Univ. de Santiago de Compostela); C. U. Moulines (Univ. München); C. Moya (Univ. de Valencia); L. Peña (C.S.I.C., Madrid); F. Recanati (C.R.E.A., Paris); B. Smith (Univ. of Buffalo); A. Sobrino (Univ. de Santiago de Compostela); C. Thiebaut (Univ. Carlos III de Madrid); J. Vázquez (Univ. de Santiago de Compostela); L. Villegas (Univ. de Santiago de Compostela); T. Williamson (Univ. of Edinburgh).

Organising committee / Comité organizador:

José L. Falguera (Univ. de Santiago de Compostela) [President / Presidente]
M^a Uxia Rivas (Univ. de Santiago de Compostela) [Secretary / Secretaria]
José M. Sagüillo (Univ. de Santiago de Compostela) [Secretary / Secretario]
L. Fernández-Moreno (Univ. Complutense de Madrid)
M. García-Carpintero (Univ. de Barcelona)
C. Moya (Univ. de Valencia)

Lectures / Ponencias:

M. García-Carpintero (Univ. de Barcelona)

Token-reflexivity and indirect discourse

R. Giere (Univ. of Minnesota)

Scientific cognition as distributed cognition

C. Martínez Vidal (Univ. de Santiago de Compostela)

The intrinsic logical status of second order logic

C. Moya (Univ. de Valencia)

Emociones, racionalidad y responsabilidad

L. Peña (C.S.I.C., Madrid)

Dos principios de permisión: algunos problemas pendientes en la lógica jurídica

F. Recanati (CREA, Paris)

Quotation

B. Smith (Univ. of Buffalo)

On analytic vs. continental philosophy. A case study in the ontology of ideas and institutions

C. Thiebaut (Univ. Carlos III de Madrid)

Tolerancia negativa y positiva: actitudes reflexivas en el espacio público

T. Williamson (Univ. of Edinburgh)

Scepticism and evidence



INDEX / ÍNDICE

Preface.....	13
Prefacio.....	15
1. PHILOSOPHICAL LOGIC / LÓGICA FILOSÓFICA.....	17
ALONSO, E.: <i>The Kleene's homework (a historical revision of the origins of computation)</i>	19
ARRIETA URTIZBEREA, A.: <i>Relatividad ontológica e identidad relativa y absoluta</i>	29
BARRY, E. & LAZZER, S.: <i>Contrafactividad y derrotabilidad</i>	41
HERRERA GONZÁLEZ, J. R. & VÁZQUEZ CAMPOS, M.: <i>Combinaciones de lógica temporal y epistémica</i>	55
IACONA, A.: <i>The usual argument for propositions</i>	65
MARRAUD, H. J.: <i>Sorites paradox from a logical point of view</i>	79
MARTÍNEZ FERNÁNDEZ, J.: <i>Sobre la relación entre la paradoja de Russell y la del mentiroso</i>	87
OLLER, C. A.: <i>Psicologismo y normativismo en lógica: El caso del razonamiento revocable</i>	95
PÉREZ OTERO, M.: <i>Elección de constantes lógicas y elección de postulados semánticos</i>	101
RAHMAN, S. & ROETTI, J. A.: <i>Dual intuitionistic paraconsistency without ontological commitments</i>	111
SALTO, F.: <i>Finitud, verdad y vaguedad</i>	123
TATIEVSKAIA, E.: <i>On logical form</i>	133
VILANOVA ARIAS, J.: <i>El gato está sobre la alfombra pero yo no lo creo</i>	143
2. ONTOLOGY / ONTOLOGÍA.....	155
BARES PARTAL, M.: <i>El realismo metafísico y la necesidad de renovar la filosofía</i>	157
D'AGOSTINI, F.: <i>There is no truth - there might be nothing: Elenchos of meta-theoretical negative statements and ontological factualism</i>	169
ELDER, C. L.: <i>Physicalism and the fallacy of composition</i>	181
LIZ, M.: <i>Las fronteras de lo físico</i>	193
PAGÈS, J.: <i>Laws, dependence and formal relations</i>	203
PARISOLI, L.: <i>The ontology of norms and the decline of the analytical legal positivism</i>	211

ROMERALES, E.: <i>The necessity of ignorance and some puzzles concerning identity and individuation</i>	219
SCHNEIDER, C.: <i>Two proposals for formalizing a bundle theory based on universals</i>	229
VICENTE, A.: <i>Vertical dependencies and the exclusion problem</i>	239

3. EPISTEMOLOGY AND PHILOSOPHY OF SCIENCE /

EPISTEMOLOGÍA Y FILOSOFÍA DE LA CIENCIA	251
ALCOLEA-BANEGAS, J. & LORENTE-TALLADA, J. M. & ÚBEDA-RIBES, J. P.: <i>Fallibilism and realism in Lakatos</i>	253
BAUMANN, P.: <i>Knowledge and relativity</i>	265
BENGOETXEA, J. B.: <i>Representación implícita y nada que representar: Crítica al Der Raum de Carnap</i>	273
BOZIN, D.: <i>Limits in science</i>	285
BRONCANO, F.: <i>La interpretación de las estructuras sociales (de los otros) y el individualismo metodológico</i>	291
CANCELA SILVA, M. C.: <i>Teoría causal de la referencia y realismo interno</i>	299
ETXEBERRIA, A.: <i>Problemas de la definición de función biológica</i>	305
GARCÍA ENCINAS, M. J.: <i>Seis o siete argumentos en pro del principio de causalidad</i>	317
GUERRERO DEL AMO, J. A.: <i>Realismo y relativismo en la filosofía de Putnam</i>	331
GUERRERO PINO, G.: <i>Kuhn y el problema de la elección racional de teoría</i>	343
IRANZO, V.: <i>Verdad y éxito empírico</i>	355
MISSELHORN, C.: <i>The turn towards a priori justification</i>	367
MORENO, A.: <i>Nuevas tendencias en ciencias cognitivas y filosofía naturalizada de la ciencia</i>	377
RODRÍGUEZ ALCÁZAR, F. J.: <i>Naturalismo en filosofía de la ciencia: pasado y futuro</i>	389
RODRÍGUEZ FERNÁNDEZ, P.: <i>Incommensurabilidad léxica y especialización en el último Kuhn</i>	403
TXAPARTEGI, E.: <i>Constructivismo sin empirismo: crítica a la teoría sensualista de van Fraassen</i>	413
VEGA ENCABO, J.: <i>El "principio de familiaridad" ¿existe un argumento russelliano a su favor?</i>	423

4. PHILOSOPHY OF LANGUAGE / FILOSOFÍA DEL

LENGUAJE	435
BLANCO SALGUEIRO, A.: <i>El holismo como mereología</i>	437
CORREDOR, C.: <i>Significado: privado, público, compartido</i>	447
DOMINGO BELANDO, A.: <i>Presuposición e implicatura convencional</i>	457
FERNÁNDEZ LANZA S. & SOBRINO CERDEIRIÑA, A.: <i>Algunos aspectos filosóficos y computacionales de la sinonimia</i>	469
FERNÁNDEZ MORENO, L.: <i>Los términos de género natural como designadores rígidos</i>	483
JALES RIBEIRO, H.: <i>Towards a reformulation of the analytical concept of the history of philosophy</i>	493
MACIÀ FÀBREGA, J.: <i>Naming and coordination</i>	501
PÉREZ-ILZARBE, P.: <i>El análisis russelliano de las expresiones denotativas y la teoría medieval del descenso</i>	511

QUESADA, D.: <i>Frege and Heidegger on reference determination and sense</i>	525
RIVAS MONROY, M. U.: <i>La filosofía analítica y la tradición analítica: Un siglo a vueltas con el significado</i>	535
ROSALES GALIÑANES, D.: <i>El papel del lector en semiótica</i>	547
VALOR ABAD, J.: <i>Davidson y Tarski, ¿podemos reducir significado a verdad?</i>	559
VICARIO ARJONA, I.: <i>Sobre la comprensión de los nombres propios</i>	569
ZILHÃO, A.: <i>A radical interpreter's paradox</i>	581
 5. PHILOSOPHY OF MIND / FILOSOFÍA DE LA MENTE	 593
BORDES SOLANAS, M.: <i>Irracionalidad motivada: El caso del autoengaño</i>	595
FERNÁNDEZ PRAT, O.: <i>Perceptual consciousness and the reflexive character of attention</i>	607
LAWLER, D.: <i>Kant on the spontaneity of the mind and the nature of the self</i>	617
MARTÍNEZ MANRIQUE, F.: <i>Contenido conceptual y no conceptual: ¿Autonomía o interdependencia?</i>	633
RODRÍGUEZ GONZÁLEZ, M.: <i>Shoemaker y el problema de los qualia</i>	647
SCHRÖDER, J.: <i>Mental causation: The supervenience argument and the proportionality constraint</i>	659
 6. PRACTICAL PHILOSOPHY / FILOSOFÍA PRÁCTICA	 673
AUSÍN, F. J. & PEÑA, L.: <i>Los derechos positivos desde una perspectiva analítica</i>	675
BETZLER, M.: <i>Rational deliberation and reactive attitudes</i>	687
MARZANO-PARISOLI, M. M.: <i>A plea for a new analytical interpretation of G. E. Moore's ethical philosophy</i>	697



Preface

Prefacio





PREFACE

This volume includes the majority of the contributions presented in the International Congress of *Analytic Philosophy at the Turn of the Millennium*, held in Santiago de Compostela, from the 1st to 4th of December 1999, which was organised by the Department of Logic and Philosophy of Science, University of Santiago de Compostela, and by the Spanish Society of Analytic Philosophy (SEFA), and aided by the European Society for Analytic Philosophy (ESAP).

The structure of this volume reflects the sections to which it was possible to submit contributions to the congress, and which, to our understanding, make up the thematic corpus of the tradition that is rooted in what has been called Analytic Philosophy. These sections are: Philosophical Logic, Ontology, Epistemology and Philosophy of Science, Philosophy of Language, Philosophy of Mind and Practical Philosophy.

Analytic Philosophy is one of the most influential traditions of thought in the contemporary field of philosophy, and the turn of the Millennium seemed to us to be a symbolic occasion for confirming its current state and its future prospects in the different areas. The origins of this tradition can be found in the works of thinkers such as Frege, Russell, Wittgenstein, Austin, Ryle, Strawson, Schlick, Carnap, Reichenbach, Tarski, amongst many others we could name, who, between the end of the Nineteenth Century and the middle of the Twentieth Century, paved the way for the so-called linguistic turn in philosophy. Some of the characteristic criteria that were established in these origins appear today to be subjected to criticism and revision. The consideration of social, moral, historical and economic factors, amongst others, is a permanent feature, and not simply a transitory one in the reflection of the heirs to the analytic school. The contributions that are included herein are a sample of the degree in which those original characteristic criteria of Analytic Philosophy have survived or are being modified. Readers of this volume will have the opportunity to verify the current state of the matter for themselves, on the basis of the papers included.

The congress was organised with the series of cultural activities which, in celebration of 1999, the Holy Year in Santiago de Compostela, have been

taking place in Santiago and other towns and cities of Galicia (Spain). The ancient tradition of the pilgrimage to Compostela has been a contributing factor to the academic enrichment of its University, due to which people and ideas from the intellectual field have arrived from all over the world. The aim of organising this congress was precisely for those interested in Analytic Philosophy and in the ideas developing from the particular style of this tradition, from all corners of the globe, to meet in the capital of Galicia, and more precisely, in its five-hundred-year-old University.

We have recently witnessed the constitution of the ESAP in Europe and, later, the SEFA in Spain with the intention of promoting the style of Analytic Philosophy. These societies assumed and made theirs the proposal for the congress, which had initially arisen from amongst the members of the Department of Logic and Philosophy of Science in the University of Santiago. The hope of all of us who have worked in this project is that the congress and this publication should contribute to cultivating the analytic style of thought.

The International Congress of *Analytic Philosophy at the Turn of the Millennium* has been made possible thanks to the efforts of many. Working with and alongside the editors of this work were Carlos Moya Espí, Luis Fernández Moreno and Manuel García-Carpintero in the organising committee. The members of the scientific committee made the selection of the proposed papers possible. We are extremely grateful for their collaboration, especially considering that it was only possible to invite a small number of them to be speakers at the congress. Everyone who contributed in one manner or another were the protagonists that made it possible to achieve our original objectives. We would like to give special thanks, for their dedication and enthusiasm shown on the organisational side and in the publication of this volume, to María Caamaño Alegre, Celeste Cancela Silva and Patricia Rodríguez Fernández, with special mention for Santiago Fernández Lanza and Diego Rosales Galifianes. Finally, we would also like to express our gratitude to the sponsoring bodies, without whose financial aid the congress and this publication would not have been realised. The sponsors were: Xacobeo'99; Xunta de Galicia-Consellería de Cultura, Comunicación Social e Turismo; Ministerio de Educación y Cultura (Secretaría de Estado de Universidades, Investigación y Desarrollo); Universidade de Santiago de Compostela; Xunta de Galicia-Consellería da Presidencia (Secretaría Xeral de Investigación e Desenvolvemento); y Residencia Universitaria San Martín Pinario.

Santiago de Compostela, November 1999

J. L. Falguera, M^a. U. Rivas y J. M. Sagüillo

PREFACIO

Este volumen recoge la mayoría de las contribuciones presentadas en el Congreso Internacional *La Filosofía Analítica en el Cambio de Milenio*, celebrado en Santiago de Compostela del 1 al 4 de Diciembre de 1999, que fue organizado por el Área de Lógica y Filosofía de la Ciencia de la Universidad de Santiago de Compostela y por la Sociedad Española de Filosofía Analítica (SEFA) y bajo los auspicios de la European Society for Analytic Philosophy (ESAP).

La estructura de este volumen refleja las secciones a las que se pudieron presentar aportaciones al congreso y que, a nuestro entender, constituyen el *corpus* temático de la tradición que hunde sus raíces en lo que se ha venido denominando Filosofía Analítica. Estas secciones son: Lógica Filosófica, Ontología, Epistemología y Filosofía de la Ciencia, Filosofía del Lenguaje, Filosofía de la Mente y Filosofía Práctica.

La Filosofía Analítica es una de las tradiciones de pensamiento más influyentes en el escenario filosófico contemporáneo. El cambio de milenio se presentaba como una ocasión simbólica para constatar el estado actual y las perspectivas de futuro de la misma en sus diferentes ámbitos. Podemos situar los orígenes de la mencionada tradición en la obra de pensadores como Frege, Russell, Wittgenstein, Austin, Ryle, Strawson, Schlick, Carnap, Reichenbach, Tarski..., y otros muchos que, entre finales del siglo XIX y la primera mitad del siglo XX, propiciaron lo que se conoce como el giro lingüístico en filosofía. Algunos de los criterios característicos asentados en tales orígenes parecen estar sometidos actualmente a crítica y revisión. La consideración de factores sociales, morales, históricos, económicos, etc. constituye un rasgo permanente y no simplemente transitorio en la reflexión de los herederos de la corriente analítica. Las contribuciones aquí recogidas son una muestra de en qué medida aquellos criterios originales característicos de la Filosofía Analítica perduran o están siendo modificados. El lector de este volumen tiene la oportunidad de constatar por sí mismo, a partir de los trabajos aquí recogidos, el estado de la cuestión.

El congreso se ha organizado dentro del conjunto de actividades culturales que, con ocasión de la celebración del año santo compostelano de

1999, han tenido lugar en Santiago de Compostela y otras poblaciones de Galicia (España). La vieja tradición de peregrinaje a la ciudad compostelana ha contribuido al enriquecimiento académico de su universidad gracias a que con ella han llegado a Compostela personas e ideas provenientes de otros centros del quehacer intelectual. La pretensión al organizar este congreso era precisamente que personas interesadas en la filosofía analítica e ideas gestadas en el estilo peculiar de esa tradición, procedentes de diferentes lugares del mundo, se encontraran en la capital de Galicia y más concretamente en el seno de su cinco veces centenaria universidad.

Recientemente se han constituido la ESAP en Europa y, posteriormente, la SEFA en España con el ánimo de fomentar el estilo de la filosofía analítica. Estas sociedades asumieron e hicieron suya la propuesta de congreso que inicialmente surgió de entre los miembros del área de lógica y filosofía de la ciencia de la universidad compostelana. La confianza de todos los que hemos trabajado en este proyecto es que el congreso y esta publicación contribuyan al cultivo del estilo de pensamiento analítico.

El Congreso Internacional *La Filosofía Analítica en el Cambio de Milenio* ha podido realizarse gracias al esfuerzo de muchos. Junto con los que editamos estas actas han estado en el comité organizador Carlos Moya Espí, Luis Fernández Moreno y Manuel García-Carpintero. Los miembros del comité científico han hecho posible la selección de propuestas de comunicaciones. A su trabajo generoso no podemos más que estar agradecidos, máxime cuando sólo unos pocos de ellos han podido ser invitados como ponentes al congreso. Todos los que han participado de una u otra manera en el congreso son los protagonistas que hacen posible el logro de los fines propuestos. Queremos resaltar la dedicación y el entusiasmo, que en aspectos organizativos y para la edición de este volumen, han mostrado María Caamaño Alegre, Celeste Cancela Silva, Patricia Rodríguez Fernández y muy especialmente Santiago Fernández Lanza y Diego Rosales Galiñanes. Finalmente, expresamos nuestra gratitud también a las entidades patrocinadoras sin cuya financiación la realización del congreso y la publicación de estas actas no hubiera sido posible. Estas entidades patrocinadoras han sido: Xacobeo'99; Xunta de Galicia-Consellería de Cultura, Comunicación Social e Turismo; Ministerio de Educación y Cultura (Secretaría de Estado de Universidades, Investigación y Desarrollo); Universidade de Santiago de Compostela; Xunta de Galicia-Consellería da Presidencia (Secretaría Xeral de Investigación e Desenvolvemento); y Residencia Universitaria San Martín Pinario.

Santiago de Compostela, Noviembre de 1999

J. L. Falguera, M^a. U. Rivas y J. M. Sagüillo

Philosophical Logic

Lógica Filosófica





THE KLEENE'S HOMEWORK

(A historical revision of the origins of computation)

ENRIQUE ALONSO
Universidad Autónoma de Madrid

ABSTRACT: Diagonalization technique threatens any attempt to describe by finitary means a class of mathematical objects. Nevertheless, the set of effective functions -i.e. the set of functions whose values can be obtained following some algorithmic process- does not diagonalize. Diagonalization does not lead to any new effective function not yet included in the class. Some years later, Gödel admits that this aspect of the notion of effectiveness seems to respond to some kind of miracle which deserves special attention. In this paper I try to examine the strategy followed by Kleene and Church to assert that effectiveness resists diagonalization. A consequence of this analysis is a better understanding of Church's Thesis.

KEYWORDS: Church's Thesis, computability, foundations of logic.

1. What's wrong?

Church's Thesis -*Ct* now on- is one of the most important statements in theoretical computer science. It can be stated as follows -there is a lot of equivalent alternatives:

[1] Every effective numerical function can be expressed in terms of some computable (recursive, λ -definable, etc) function.

It has been considered one of these rare mathematical statements which do not constitute properly axioms nor theorems of some more comprehensive theory, but empirical hypothesis only testable by experience. Church states his thesis for the first time during the period starting at the fall of 1933 and finishing at the spring of 1934. Kleene, a Church's student in the years we are talking about, describes his opinion with respect to the *Ct* in a way which results strange from a contemporary point of view. His comments offer the starting point for my analysis.

[Kleene 1981]: "When Church proposed this thesis, I sat down to disprove it by diagonalizing out of the class of the λ -definable functions. But, quickly realizing that the diagonalization cannot be

done effectively, I became overnight a supporter of the thesis." [Kleene 1981, p.59].

What's wrong in this statement? What did Kleene realize this night while he tried to do his *homework*? The first thing worth mentioning is the impossibility to conclude the diagonal procedure over the class just considered. When do we realize that a diagonal argument cannot be done effectively? The second aspect which results a bit puzzling from our present understanding of the matter is the consequences it has with respect to *Ct*. How can this fact give any support to *Ct*?

The second aspect seems to be closely related to a defective understanding of the logical structure of a diagonal argument. The reason we can have to find puzzling the statement in which Kleene declares the impossibility to conclude a diagonal argument has something to do with the influence of the calculus of natural deduction for elementary logic over our understanding of the structure of a diagonal argument. Natural deduction is not a decision procedure for any logic whatsoever. This feature does not depend on the logic itself, but on the definition of this procedure. If a deduction finishes, the formula which stand in the last line becomes a theorem of the logic in question. But, what if the deduction does not finish? Everybody knows that this very question is a nonsense. There is no way to determine if a deduction will never conclude -given a formula fixed as target- or if it will do some steps later. The way we use to consider diagonalization constructions seems quite similar to this general model.

I don't mean either that diagonalization procedures should be considered as genuine algorithms. In fact they are not. What I affirm is that Church and Kleene discovered a way to work with diagonalization so that the impossibility to diagonalize on some class can be stated in the same sense we probe genuine mathematical theorems. Once we establish this thesis -the main thesis in this issue- we understand the reasons Kleene had to believe in *Ct*.

2. The logical structure of diagonalization.

Diagonalization is a mathematical tool well known to everybody. Cantor proof relative to the non enumerability of the reals, and many other results in which diagonalization plays a fundamental role are an important component of the core of basic facts that logicians, philosophers and historians of logic and mathematics should not ignore. Nevertheless, it is not to say that everybody agrees with the form these proofs have. I don't think either we agree on the very description of these constructions. In fact, it is not a difficult task to find more than a way to describe a diagonal construction amongst the many which recent history offers. This is owed to the partial ignorance of the precise logical structure of actual diagonalization constructions. In fact, I think that we can find several varieties of diagonalization whose distinctive aspects are clear enough as to consider as

objects belonging to a large class, perhaps an entire family. In what follows I offer two constructions closely related. Their logical structure can be described by means of basic logical tools, and both play a role in the process which leads Church and Kleene to formulate Ct .

In what follows I am going to make use of some special symbols useful to establish the format of the variants of diagonalization I have mentioned. Let be C some class of mathematical objects given by a definition not depending on the diagonal techniques of which we are going to make use. Now let be M a matrix whose elements correspond to the members of C . M tries to represent the class C assigning to every point in the matrix a value in C which stands for this point. Once we have a definition of the matrix M a new object is obtained by a mere enumeration of the points along the diagonal arrow of this matrix. This is the *diagonal object* we represent as d . From d we define the *antidiagonal object* d^* taking d and modifying its definition so that it differs from d in every point. It is immediate to realize that d^* can be obtained from d by means of the most elementary operations, so that the consequences that derive from the existence of d^* can be derived also from the existence and definition of d itself. Now we find two objects f , and f^* which translates d , and d^* in terms of the proper tools used by C . It does not mean that f , and f^* are *in fact* objects in C . That depends on the construction of M and C and, as we will see, of extralogical considerations. Finally, by $C \simeq M$ I mean that M is some sound and complete representation of the class C .

Now it is possible to depict two different schemata corresponding to different diagonalization constructions,

Schema I.

1. Definition of the class C .
2. Definition of the matrix M from de definition of C .
3. **Hypothesis:** $C \simeq M$
4. Definition of d [from M]
5. Definition of d^*
6. If $d \in M$ then $d^* \in M$ [elementary logic]
7. $d^* \notin M$ [diagonalization]
8. $d \notin M$
9. $f =_{\text{def}} d$
10. $f \in C$ (!) [from extralogical considerations and 9]
11. $f \notin M$ [alternatively $f^* \notin M$]
12. $C \not\simeq M$ [by 10 and 11]
13. $C \not\simeq M$ [reductio 3-12]

Schema II.

1. Definition of the class C .
2. Definition of the matrix M from de definition of C .
3. **Hypothesis:** $f \in C$
4. Definition of d [from M]
5. Definition of d^*
6. If $d \in M$ then $d^* \in M$ [elementary logic]
7. $d^* \notin M$ [diagonalization]
8. $d \notin M$
9. $f =_{\text{def}} d$
10. $C \simeq M$ (!) [from extralogical considerations and 9]
11. $f \notin M$ [alternatively $f^* \notin M$]
12. $f \notin C$ [by 10 and 11]
13. $f \notin M$ [reductio 3-12]

Both schemata have well known instances. Diagonalization proofs along the first schema are the demonstration that the set of real numbers is not denumerable, or that the power set of any infinitely numerable set is not denumerable, etc. We make use of the second one when we probe any current version of the *halting problem*.

If we consider the logical structure of these diagonalization schemata it is possible to understand some of the reasons Church and Kleene had to say that diagonalization does not affect to the acceptability of Ct . If we take the schema I we see that the entire proof depends on the acceptability of the proposition listed in the step 10. To admit that $f \in C$ amounts to say that the definition of f in terms makes of f a genuine object of C . This situation, which is easy to admit in most cases, can be dubious in many others. Anyway I don't think it is a matter of pure logic, but a combination of logic, mathematics, a philosophical conceptions working together. When the proof arrives at the point in which it corresponds to analyse if f belongs to C all these considerations act to give an answer to this question. If we *decide* that f is a proper object in C , diagonalization finishes showing that M does not represent C . If, on the other hand, f is not accepted as a genuine object of C we should give up the demonstration in that point. The result of this incomplete proof is the reinforcement of the statement which serves as initial hypothesis.

As I will explain later this is only one part of the reasons Kleene found to talk about Ct in the way he does in [Kleene 1981]. In fact what they -Kleene and Church- discovered in the time we are talking about is something considerably stronger. They found a way to combine the schema I with the schema II to probe that diagonalization cannot effectively done. Nevertheless, the main intuition on which rests the work done by Kleene depends on the kind of failure I have exposed above.

3. Mathematical Inquiry at stake. Imaginary solutions, real solutions.

When Church proposed his system for foundations of arithmetic mathematical inquiry was not the peaceful land we are used to live in. Undecidability theorems and diagonalization techniques threaten the very foundations of mathematical building. The following text offers a vivid description of what is in the air.

[Rogers 1967]: "It is evident that the diagonalization method has wide scope, for it is applicable to any case where the sets of instructions ... can be effectively (i.e. algorithmically) listed. At first glance, it is difficult to see how a formal characterisation can avoid such effective listing and still be useful. [...] It suggest the possibility that no single formally representable class of algorithmic functions can correspond exactly to the informal notion of algorithmic function. Perhaps, no

matter what...symbolism and...specifications we choose, that symbolism and those specifications can be augmented by stronger symbolism and more complex "effective" operations to yield new functions...Perhaps, indeed, the algorithmic functions form a non-denumerable class, and perhaps there is a spectrum of algorithmic computability upon which all functions fall." ([Rogers 1967, p.11]). [I quote following Webb 1980, p.178].

Nevertheless Church does not think that everything is lost. This the impression one obtains when reading the following text:

[Church 34]: "Given a system of symbolic logic, let us try to construct the function of positive integers such that there is no formula in the system that stands for it. What we must do is first to enumerate all formulas, and then, going through this enumeration, to pick out in order those formulas which stand for functions of positive integers. The result is an enumeration of all formulas which stand for functions of positive integers. And if we let $f_n(x)$ be the function of positive integers represented by the n th formula in this enumeration, then $1+f_x(x)$ is the function of positive integers such that there is no formula in the system that stands for it. But this function $1+f_x(x)$ is not, in general, defined in such a way that it is always possible to calculate its value for a given positive integer x . For, in the process of going through the list of all formulas and picking out those which stand for functions of positive integers, we may at some stage find a formula about which we do not know whether or not it stands for a function of positive integers.

...Indeed, to be sure of always being able to determine whether a given formula stands for a function of positive integers, we must have discovered a method of procedure which would enable us to solve any problem of number theory whatever. Therefore the infinite sequence (about which we have been talking) of all formulas which stand for functions of positive integers almost certainly is not such an infinite sequence that it is possible to calculate as many terms of it as we please. And therefore the function $1+f_x(x)$ has not been defined in a way which could be called constructive, but has merely been proved by an indirect argument to exist." [Davis 1982, p.9]

Church shows a great ability when examines what seems to be the *possibility conditions* that should meet any complete description of the class of effective functions. He avoids to say anything he does not know, so that he refuses to speak directly of the existence of such a complete description - λ -calculus. Anyhow he imagines the way things should be if such a complete description actually were to exist. It is not a difficult task to observe that the main argument rests on an *imaginary* proof of the type I have depicted in *schema I*. Church is not talking about an object language construction able to

count as a real proof. Instead, he formulates a metatheoretic proof which fixes the main characteristics a real proof should have. Were as it were, it remains to confirm that the indirect existence of the antidiagonal object - $1+f_x(x)$, in this case- is not to be transformed some steps later in a real argument about the existence of this object. We need to complete the list of possibility conditions with a proof that the antidiagonal object does not exist. This part of the argument is going to be supplied by Kleene's work about undecidability.

The diagonal construction conjectured by Church is metatheoretical and imaginary in character. It affects to a wide range of classes of mathematical objects. In fact, any class such that:

- [2] i. Its vocabulary is denumerable, and
- ii. Has an inductive definition,

lies under the scope of Church's argument.

If we go back to the domain of effective functions, the argument Church offers in this text can be resumed as follows:

- [3] i. If the class of effective functions is going to have a complete description, then it cannot exist any procedure being able to determine in an effective way whether a function is defined for every element in its range.
- ii. If there exists a procedure to perform this task, then there is a solution to any question we can ask for in number theory.

Church was right about the consequences of his position but he does not offer a conclusive proof relative to the non existence of the function mentioned in [3.i]. I think that Church only saw one half of the entire proof and that was Kleene who discovered the way to confirm Church's conjecture reaching the result Church himself had predicted. When Kleene realized that the metatheoretical diagonal argument formulated by Church could be enlarged by a *second* diagonal argument as to yield a conclusive proof, he became overnight a supporter of the thesis -CT.

The part added by Kleene is not less imaginary than the Church's construction. The difference consists in its character. This time the proof takes place in the object language. Kleene's diagonalization could be summed up as follows:

- [4] i. Let $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_i, \dots$ be an effective and complete enumeration of the set of well formed formulas in λ -calculus.
- ii. Let $\theta(x) = 1$ if ε_x reduces to some numeral n .
 0 otherwise.
- iii. Now, we have to admit that $\theta(x)$ does not exist for otherwise $1+\theta(x)$ would also be λ -definable contradicting the existence of the enumeration presented above.

Let us note that the class of objects which are enumerated in [4.1] is not the class of effective functions, nor the class of λ -definable functions, but a set of expressions that meets conditions listed in [2]. This second diagonal argument is a classical undecidability result which imposes some conditions of the objects is referred to. In fact, it establishes that $\theta(x)$ is not an object in the class. Nevertheless, it is not a real proof. It is an imaginary solution which depends on a list of results not yet probed -at least not published as such. To accept the required generality of this imaginary diagonal argument requires the introduction of Kleene's minimalization operator μ , the Kleene's Enumeration Theorem (Normal Form Theorem), and the equivalence between λ -definable functions, general recursive functions and functions expressible in the formalism of primitive recursive functions plus minimalization rule. In fact I maintain that these results are obtained by Kleene as a part of the program now depicted. All of them are required if the class of effective functions has to resist to diagonal techniques, i.e., if this class of functions has to satisfy the statement of *Ct*. To put it otherwise, we require this kind of technical steps to make that the entire imaginary process I have exposed can be transformed in a real demonstration able to give a positive support to *Ct*. So what I maintain can be summarised as follows:

[5] *Ct* was established in full detail before the results which make it a plausible conjecture were probed. In fact, *Ct* is responsible for the research which lead to many of the basic theorems in theory of computation.

I cannot give definitive reason to defend that was Kleene and not Church who discovered for the first time the strategy I have been talking about lines above. The combination of two diagonal arguments, one which follows the schema I, and another one which follows the schema II, make it possible to conclude, with the force of a mathematical demonstration, that the diagonalization does not lead outside of the intended class. Of course this is not to say that *Ct* should be considered as any other theorem in mathematics. Its acceptability still depends on the existence of some other formalism not equivalent to any other item in the list of systems described till now. It is possible that both Kleene and Church were conscientious of the entire story. Anyhow is Kleene the one who describes the things such as they probably were during this marvellous period in which everything seemed to happen.

[Kleene 36.1]: "In §2, the problem is raised, which systems of equations define recursive functions under the general definition. The systems which do cannot be recursively enumerated, if by a recursive enumeration is understood one such that the numbers ordered by the Gödel method to the systems of equations in the enumeration are a recursive sequence (i.e. the successive values of a recursive function of one variable), since from any recursive sequence of such numbers

we can obtain the recursive definition of a new function by the familiar process of diagonalizing and adding 1. For the same reason, a recursive process of deciding which systems define recursive functions is unattainable, if by a recursive process is meant one such that there is a recursive function of the corresponding numbers whose value is 0 or 1 according to the result obtained". [Kleene 1936, p.238]. And lines below we can read also,

[Kleene 36.2]: "The definition of general recursive function offers no constructive process for determining when a recursive function is defined. This must be the case, if the definition is to be adequate since otherwise still more general "recursive" functions could be obtained by the diagonal process". [Kleene 1936, p.248].

Some years later, in 1938, Kleene offers what now can be considered the standard model in relation to effectiveness. Total functions are given up in favour of partial functions. This decision offers a second solution to the diagonal argument I have describe from Church's cite. In this case the double diagonalization argument collapse into an unique diagonal construction allowing a partial behaviour for the universal function which corresponds to d in the schema I. I am not interested in this solution because it was not what Church, Kleene, Gödel an many others taken into account to convince themselves of the good reassons to accept *Ct*. They -but Gödel- were supporters of the thesis even when total functions were the only kind of numerical functions mathematicians had in mind. Therefore it is not difficult to understand the surprise Gödel manifest in relation to the ability of our notion of effectiveness to resist diagonalization where concepts closely related are not able to resist.

[Davis 1982; from the conference lectured by Gödel in the Princeton University Conference 1965]: "In all other cases treated previously, such as demonstrability or definability, one has been able to define them only relative to a given language, and for each individual language it is clear that the one thus obtained is not the one looked for. For the concept of computability however, although it is merely a special kind of demonstrability or decidability the situation is different. By a kind of miracle it is not necessary to distinguish orders, and the diagonal procedure does not lead outside the defined notion". [Davis 1982, p.16].

REFERENCES

- Alonso, E: "Ingenio e Industria", *Theoria*, vol. 14, nº35, mayo 1999.
 Church, A: "The Richard Paradox", *Amer. Math. Monthly*, 41, 1934, pp.356-361.
 Davis, M: *The Undecidable. Basic Papers on undecidable propositions, unsolvable problems, and computable functions*, Raven Press, N. York, 1965.
 Davis, M: "Why Gödel Didn't Have Church's Thesis", *Information and Control*, 54, 1982. pp.3-24.

Kleene, S.: "General recursive functions on natural numbers", *Mathematische Annalen*, 112, pp.727-742. Compiled in Davis, 1965, pp. 237-254.

Kleene, S.: "Origins of Recursion Theory", *Annals of the History of Computing* vol. 3, n° 1, 1981. pp. 52-67.

Enrique Alonso

U.A.M.

e-mail: Enrique.alonso@uam.es





RELATIVIDAD ONTOLÓGICA E IDENTIDAD RELATIVA Y ABSOLUTA

AGUSTÍN ARRIETA URTIZBEREA
UPV-EHU

ABSTRACT: One of the most controversial issues concerning identity is whether identity is relative or absolute. D. Wiggins has championed the defence of the absolutist conception. The opposite view, say, the relativistic one, was advanced by P. Geach. Noonan argues, on the basis of some proposals put forward by M. Dummett, that “the relativist position is nearer to being correct”, even though is not wholly correct (Noonan (1996), 25-31). Against Noonan I want to incline the discussion on the absolutist side.

KEYWORDS: Absolute identity, relative identity, equivalence relation, criterion of identity, ontological relativity, countable general terms.

1. Introducción

Una de las grandes discusiones acerca de la identidad gira en torno a la naturaleza absoluta o relativa de la misma. Noonan (en Noonan (1996)) viene a señalar que ninguna de las dos perspectivas es completamente correcta aunque, desde su punto de vista, la posición relativista goza de cierta ventaja. En este trabajo mi objetivo es presentar los fundamentos de la posición relativista para concluir, contrariamente a Noonan, que en los marcos en los que la discusión se ha planteado es el punto de vista absolutista el que está en una situación favorable. Por otro lado, teniendo en cuenta el estrecho vínculo existente entre identidad y ontología, cabría pensar que hay una relación de dependencia obvia entre la perspectiva relativista en el debate acerca de la identidad y la tesis de la relatividad ontológica. Quine ha argumentado en contra de esta hipótesis al mantenerse en una posición absolutista con respecto a la identidad a pesar de su relativismo ontológico. En cualquier caso, algo que se desprende de este trabajo es que las concepciones ontosemánticas y del lenguaje que el partidario de la identidad relativa arrastra poco tienen que ver con la tesis del relativismo ontológico. Espero que la descripción que aquí se desarrolla contribuya a hacer más plausible y más gráfico el contraste entre los dos relativismos.

El orden de exposición es el mismo que Dummett sigue en un conocido trabajo (Dummett (1991)) sobre la idea de identidad de Geach. En primer lugar se plantea la discusión en el contexto de los lenguajes formales, para a continuación pasar a la discusión en términos generales. Posteriormente se muestra que hay una clara dependencia entre los diferentes aspectos desarrollados en los dos contextos. Finalmente tengo que señalar que a partir de ahora utilizaré el término ‘el relativista’ para referirme al defensor de aquellas tesis geachianas que están en el núcleo de la posición relativista (estas tesis se presentan fundamentalmente en Geach (1962, 1967, 1969, 1973, 1991) y en Noonan (1996, 1997)).

2. Relativismo acerca de la identidad en el marco de los lenguajes formales de primer orden.

Sea L un lenguaje de primer orden con identidad. Sea T una teoría formulada en L . Sea I una interpretación de L que es modelo de T . El signo de identidad *puede* interpretarse en el marco de I como una relación de equivalencia definida en el dominio de I . En la posición del relativista es de vital importancia el hecho de “debilitar” la identidad hasta el punto de que una relación de equivalencia pueda hacer las veces de la identidad. De esta forma el enunciado ‘ $a=b$ ’ puede ser verdadero en I , a pesar de que los términos singulares ‘ a ’ y ‘ b ’ denoten objetos distintos del dominio de I . Esta es una situación ventajosa para el relativista. El absolutista tiene una respuesta: a partir de I se puede obtener otra interpretación I' donde el signo de identidad se interpreta ahora como una relación de equivalencia “absoluta”, es decir, donde no cabe que dos objetos *distintos* del dominio estén relacionados según la interpretación del signo de identidad. Con lo cual la situación es, por decirlo así, de empate o, cuando menos, de indecisión. L *puede* interpretarse de distintas maneras satisfaciendo los intereses tanto relativistas como absolutistas. El relativista puede tratar de justificar que es mejor la opción I que la opción I' . Para ello analiza L y sus interpretaciones desde un lenguaje L^* , lenguaje este último que se obtiene a partir de L por enriquecimiento alfabético. Cabe establecer una interpretación I^* para L^* obtenida a partir de I con los ajustes debidos. Puede que el enriquecimiento alfabético haya sido tal que ahora aquella relación de equivalencia que hacía las veces de la identidad para el caso de I y L no sirva como interpretación del signo de identidad al pasar a L^* e I^* . El relativista parece defender que en cualquier caso cabría interpretar el signo de identidad como una nueva relación de equivalencia definida *sobre el mismo dominio*. Obviamente el absolutista siempre podría *normalizar* I^* (igual que *normalizó* I para obtener I') de acuerdo con sus intereses. En este proceso, al interpretarse el signo de identidad como una relación de equivalencia, el relativista puede plantear lo que normalmente se entiende como su tesis. Por ejemplo, en L , se puede decir que a y b son

idénticos en el sentido que marque la interpretación I para el signo de identidad, a pesar de que *a* y *b* son distintos (en algún otro sentido).

Así, se entrelazan dos tesis relativistas:

(a) El signo de identidad siempre puede interpretarse como una relación de equivalencia en este proceso de enriquecimiento alfabético (que hay que entender en términos conjuntistas y no temporales).

(b) Por otro lado, en virtud de (a), siempre ocurre que *a* es el mismo que *b* en algún sentido, pero es distinto en algún otro sentido. Con lo cual la identidad es relativa.

En el contexto que nos ocupa (b) es una consecuencia de (a) y la verdad de (a) depende de que el lenguaje sea suficientemente pobre. Al pasar a un lenguaje suficientemente rico (con los predicados pertinentemente interpretados como para poder distinguir entre cada dos individuos del dominio de I) no cabe concluir que se dé (b). Sólo cabe la identidad absoluta involucrada desde el principio del proceso en la especificación del dominio en el metalenguaje: es la relación de identidad *asociada con el dominio* de la interpretación I.

En definitiva, el absolutista siempre tiene su salida. Una salida en cada fase del proceso que el relativista propone: *normalizar* la interpretación. Además también dispone de una salida global aceptando las fases y el proceso tal y como el relativista desea.

Destaco dos respuestas del relativista:

(1) Las soluciones absolutistas conllevan la aparición de una ontología superpoblada, cosa no deseable.

(2) Cabe especificar el dominio de una forma, digamos, no estándar, de tal forma que el absolutista se quede sin la salida global.

En relación a (1) el argumento del relativista puede presentarse de la siguiente forma:

Sean *L**, *I**, *L* e *I* los lenguajes e interpretaciones arriba señalados. Para el relativista lo adecuado al ver *L* desde *L** es considerar que el signo de identidad en *I* *debe interpretarse* como una relación de equivalencia definida en el dominio tanto de *I* como de *I**. El absolutista defiende que para el caso de *L* cabe interpretar el signo de identidad como una relación de identidad absoluta definida en el dominio de otra interpretación *I'*. Pero para el relativista esto último implica una superpoblación ontológica ya que en el sublenguaje interpretado $\langle L, I' \rangle$ hay compromiso ontológico con un tipo de entidades y en el lenguaje extendido (e interpretado) el compromiso es con entidades de otro tipo. El relativista parece defender que en la medida en que *L* es un sublenguaje de *L** (cabría considerar otros sublenguajes) en *L** se arrastran los compromisos ontológicos de los sublenguajes interpretados como el absolutista prefiere (cabría decir lo mismo a nivel de teorías formuladas en los respectivos lenguajes). El deseo de ahorro ontológico va en contra de esta

estrategia. ¿Para qué comprometernos con diferentes tipos de entidades, si cabe una “vía” de interpretación, la propuesta por el relativista, que conlleva automáticamente ahorro ontológico?

Un argumento distinto del ahora descrito pero vinculado con él dice: sea el lenguaje L interpretado como el absolutista quiere (esto es, mediante I’). El absolutista debe comprometerse con las entidades del dominio de I’. El proceso de normalización que nos ha llevado de I a I’ hace de los elementos de I’ clases de equivalencia. El relativista considera que es absurdo comprometerse ontológicamente con clases de equivalencia, por ejemplo, de seres humanos cuando nos basta con los seres humanos mismos.

Cabe dar respuestas de diferente tipo a estos argumentos basados en el ahorro ontológico (dejo de lado los argumentos relativos al propio concepto de compromiso ontológico y de herencia de los compromisos ontológicos aplicado al caso de un lenguaje y sus sublenguajes. Sobre este problema Dummett (1991) y Noonan (1997)):

- ¿Qué sentido tiene afirmar que en L hay compromiso con entidades que no son distinguibles en el propio lenguaje L? Si el relativista, en L, considera que se habla de seres humanos, a pesar de que en L sólo hay recursos para distinguir entre seres humanos que tengan sueldos distintos, es porque más allá del lenguaje, en el metalenguaje, se distingue entre seres humanos. Por ello, en realidad, *ser el mismo ser humano que* hace las veces de identidad absoluta implícita en el lenguaje interpretado. Si no hay esta identidad absoluta implícita, la propuesta relativista no tiene mayor sentido ya que en L no hay capacidad para distinguir distintos seres humanos que ganen el mismo sueldo, con lo que la ontología con la que se compromete es una ontología de clases de equivalencia (de seres humanos que ganan el mismo sueldo). En ambos casos la identidad absoluta es ineludible.

- Otra vía argumental en relación con la anterior, aunque sobrepesa más otros aspectos, es la siguiente. La tesis del relativismo ontológico afirma que desde un lenguaje (o desde una teoría formulada en un lenguaje) no hay manera de seleccionar una interpretación del mismo (o de la misma) que sea preferible (Quine (1969), pp.53-54). Si aceptamos esta tesis como correcta, entonces no cabe decir ‘*debe interpretarse*’, que es lo que el relativista acerca de la identidad plantea. Si el relativista lo único que pretende es decirnos que el signo de identidad *puede* interpretarse como una relación de equivalencia, entonces no veo ninguna razón para contraatacar. Efectivamente el signo de identidad *puede* interpretarse como una relación de equivalencia definida sobre un dominio D o *puede* interpretarse como una relación de identidad estricta sobre un dominio C. Considero que el relativismo ontológico nos permite dar el mismo valor a estos dos ‘*puede*’. Lo que sí parece es que, salvo propuesta de otros argumentos, el absolutista goza de ligera ventaja, ya que aun cuando se interprete el lenguaje tal y como el relativista “prefiere” hay una relación de

identidad absoluta en la especificación del dominio. A este problema se hace referencia en el punto (2) arriba señalado. El relativista, en primer lugar, presenta la idea de que sobre un único dominio caben diferentes relaciones de equivalencia que hacen la función de la relación de identidad. Lo que no es deseable, desde el punto de vista relativista, es una proliferación de dominios. Más allá de esta proliferación, el absolutista nos recuerda que la especificación de un dominio conlleva una relación de identidad asociada (el dominio se puede especificar por diferentes vías: como conjunto, mediante un sortal, etc.) Aquí es donde el relativista no tiene otra salida más que la de plantear problemas al procedimiento estándar para la especificación del dominio. Geach propone una estrategia de especificación del dominio mediante listas de nombres propios, donde cabe que un nombre propio posteriormente resulte ser un nombre compartido. Dummett ha analizado los problemas de los que esta estrategia adolece. La respuesta de Geach, hasta donde yo conozco, es irónica, pero no convincente: "more than once I have read the further explanation: I mean 'identical' as it is meant in the meta -language! I cannot see that this is anything but an incantation, a magic word..." (Geach (1991), p.298-299).

Parece ser que, por lo señalado hasta aquí, el relativista se ve abocado a entrar en una ruta que nos puede dar una idea del precio que hay que pagar por defender una perspectiva relativista en el marco de los lenguajes formales: la semántica estándar le resulta problemática al relativista. En cualquier caso, trasladémonos a un marco más general.

3. Los fundamentos del relativismo acerca de la identidad.

Presento algunas tesis fundamentales que subyacen al relativismo:

(R1) No hay una relación de equivalencia absoluta.

(R2) Los enunciados de identidad tiene, hablando en términos generales, cuatro componentes: una expresión para la identidad, dos términos singulares y un término general. La tesis básica es que dejando sin variar los tres primeros componentes, las variaciones del cuarto pueden generar, a su vez, variaciones del valor veritativo.

(R3) Los términos generales a los que se hace referencia en (R2) son términos sortales. Un sortal es el fruto de un proceso de desrelativización a partir de una relación de equivalencia (en general, me voy a referir a los términos contables que en opinión de Geach constituyen una subclase del conjunto de los términos sortales o sustantivos. Me limito a los contables para evitar problemas en torno a la clasificación geachiana).

(R4) La operación de contar está orientada no por un concepto, sino por una relación de equivalencia.

(R5) Hay que distinguir entre cuantificación restrictiva y no restrictiva.

(R6) Hay que distinguir entre nombres-de (of) y nombres-para (for).

(R2) y (R3) configuran la raíz del tronco relativista. Trataré de mostrar cómo cabe construir el edificio relativista a partir de estas dos tesis comprobando, a su vez, cómo el resto de las tesis penden de las mismas.

Frege argumenta a favor de la tesis que voy a denominar 'relatividad del contar': sólo tiene sentido contar (asignar un número) si se hace desde un concepto que implícita o explícitamente orienta la operación de contar: 300 hombres, 3 compañías,... Contamos 300 objetos *distintos* que caen bajo el concepto *hombre*, contamos 3 objetos *distintos* que caen bajo el concepto *compañía*. Cambiando el concepto, el enunciado cambia de valor veritativo: 'hay 300 compañías' es falso. La relación existente entre la operación de contar y la relación de identidad hace que el relativista considere algo casi trivial que la relación de identidad también deba de estar orientada por un concepto (dejamos de lado la posibilidad de plantear problemas a la tesis fregeana como en Alston-Bennett (1984)). El relativista va más lejos ya que pretende justificar que tras la operación de contar no se haya un concepto monádico, sino más bien una relación de equivalencia. Así pues, tanto el contar como la identidad son por sí solas incompletas, requieren de una relación de equivalencia que complete ambas, esto es, que dote a ambas de un sentido pleno. El enunciado de identidad es incompleto mientras no se complete con un término contable, de tal forma que asociado al sentido de ese término contable haya un criterio de identidad que viene determinado por una relación de equivalencia. En opinión del relativista no se trata de algo casual ya que un contable es fruto de un proceso de desrelativización a partir de una relación de equivalencia.

Dummett (en Dummett (1973, 1981, 1991)) ha mostrado convincentemente que no todo término contable es fruto de un proceso de desrelativización. En cualquier caso, al relativista quizá le basta con que algunos términos se comporten según la tesis de desrelativización. Dejando de lado este tipo de crítica a la tesis (R3) me centraré en el siguiente aspecto básico: compararé la conocida estrategia fregeana para el término 'dirección' con la estrategia relativista. Frege toma como punto de partida una relación de equivalencia R (de paralelismo) definida en un dominio D (de rectas). Posteriormente introduce un término ' F ' ('dirección') proporcionándonos un criterio de identidad asociado con el mismo. Este criterio de identidad es de segundo nivel (siguiendo la distinción entre dos tipos de criterios de identidad propuesta en Lowe (1997)). Es decir, el criterio de identidad no es de forma directa para F -s, sino que más bien está asociado a expresiones del tipo 'el F de x ', donde x es un elemento de D . Consideremos que una dirección es una clase de equivalencia. Si se pregunta ¿cuántas direcciones hay?, ¿es la dirección de x la misma que la dirección de z ?, la explicación fregeana de la génesis de 'dirección' nos proporciona instrumentos para las diferentes respuestas. Pero esto no quiere decir, como Dummett ha señalado, que la "teoría" fregeana sea

completa. Frege en los *Fundamentos* (par. 65) señala que todas las aseveraciones sobre direcciones deben de ser definidas y en estas definiciones se debe respetar la ley de Leibniz, en el sentido de que si algo se dice de la dirección de x , lo mismo cabe decirse de la dirección de y , si es que x e y son rectas paralelas. El modelo fregeano proporciona respuesta inmediata a las preguntas arriba señaladas. Por ejemplo, para responder a la primera pregunta basta con ver cuántas clases de equivalencia hay. Sin embargo, el modelo fregeano no proporciona respuesta inmediata a la pregunta, ¿estas dos direcciones se cruzan? Una respuesta respetuosa con las exigencias fregeanas sería la siguiente: la expresión relacional ‘se cruzan’ se aplica a dos direcciones d_1 y d_2 si existe una recta r_1 que pertenece a d_1 y existe una recta r_2 que pertenece a d_2 tales que r_1 y r_2 tienen un punto en común.

El modelo del relativista sigue al principio un proceso análogo (el famoso ‘surman’ de Geach), pero hay una diferencia importante y es por donde hay que hurgar ya que ahí quizás esté situada la herida relativista. Tenemos D y R como se ha señalado anteriormente. Obtenemos por desrelativización de R el contable F . Cabe responder a preguntas sobre el número de F -s o sobre si dos F -s son idénticos. Los problemas empiezan cuando nos preguntamos, por ejemplo, ¿algún F es G ? Los problemas empiezan porque de cualquier elemento de D se puede decir que es un F , según el relativista. En realidad el relativista no quiere que se pase a otro dominio (el conjunto cociente) donde otra relación de identidad y otras propiedades (son otras a pesar de que nominalmente puedan coincidir con las propiedades de las rectas) entran en juego, sino que lo que quiere señalar es que tenemos un dominio D y en D podemos considerar diferentes relaciones de equivalencia. El relativista soluciona los problemas relacionados con una correcta introducción de un término general en el lenguaje recurriendo a las tesis (R5) y (R6). Introduce nombres-para, incorpora cuantificación restringida en términos de nombres-para, etc., de tal forma que la introducción del término general F en el lenguaje resulte coherente y completa y pueda regularse su comportamiento en el contexto de enunciados como, por ejemplo, ‘algún F es G ’. En este juego lingüístico coherente cabe defender que dos elementos de D que estén R -relacionados sean distintos, valga la expresión, D -s pero el mismo F . Frege no aceptaría esta forma de presentación ya que, para el caso concreto de las rectas, dos rectas distintas no son la misma dirección por el simple hecho de que las rectas no son direcciones. Algunos ejemplos de Geach tienen la estructura general arriba señalada. Baste recordar los ejemplos relacionados con las palabras de un libro concreto (palabras-tipo o palabras-ejemplar), o los animales de un zoológico concreto (animales individuales o especies animales). Es claro que la respuesta a la pregunta por el número de animales en un zoológico depende de cómo se interprete la expresión ‘animales’. En cierto sentido la operación de contar es distinta según la relación involucrada.

Efectivamente, como el relativista señala, la cuenta depende de la relación seleccionada. Pero una vez establecida la relación involucrada, es decir, decidido el sentido de 'animal' no hay ningún problema en recurrir a un esquema de explicación fregeano que dé cuenta de la introducción de 'especie animal' sobre la base del dominio de los animales individuales. El relativista diría 'el animal individual a es el mismo animal que el animal individual b, pero son distintos animales individuales' (dejo de lado la cuestión de la gramaticalidad o nivel de corrección de esta expresión). Esto es verdad porque *ser el mismo animal que* no es la relación de identidad asociada al dominio al que pertenecen los animales individuales a y b. Si cuando decimos 'a es el mismo A que b' *ser el mismo A que* es la relación de identidad asociada con el dominio al que pertenecen a y b, entonces no cabe encontrar un B de tal forma que se cumpla la tesis relativista: a y b son el mismo A, pero distintos B-s.

Recapitulando, el relativista puede justificar la tesis relativista gracias a (R3). Es dudoso que los términos generales contables respondan a ese patrón de introducción en el lenguaje. En lo que a la operación de contar se refiere, los términos contables, tal y como el relativista los concibe, se comportan de acuerdo al patrón fregeano. A partir de ahí, debido a que el relativista no quiere nada más que un dominio de entidades, el relativista se ve obligado a efectuar una serie de maniobras que resultan chocantes desde el punto de vista del comportamiento habitual de los términos generales en el lenguaje ordinario. Volveré a las maniobras relativistas para analizar el ejemplo del gato Tibbles.

Por otro lado, lo curioso de (R3), como Noonan (en Noonan (1996)) reconoce, es que el criterio de identidad asociado al término contable 'F', es decir, *ser el mismo F que*, no está definido, como sería deseable, en el dominio de los F-s, sino más bien en el dominio D. Sólo así puede montar el relativista su estrategia. Noonan considera que un absolutista, frente al relativista, tiene ventaja con respecto a lo que acabo de señalar, pero considera que hay otra condición exigida a todo criterio de identidad que el absolutista no puede satisfacer, a saber, que la especificación del criterio de identidad para los F-s debe formar parte esencial del sentido completo del concepto contable F. Su argumentación es confusa y he de reconocer que no soy capaz de seguir la misma. En cualquier caso, cabe pensar en un caso paradigmático que de alguna forma se enfrenta a la tesis defendida por Noonan. Pensemos que el sentido completo del término 'conjunto' viene dado por la axiomática de Zermelo-Fraenkel. Parece que el axioma de extensionalidad constituye el criterio de identidad asociado al término 'conjunto' y en este axioma se nos dice cuándo dos conjuntos son idénticos en un sentido absoluto. Se trata además de un criterio de identidad de primer nivel. Además este criterio es un axioma entre otros, es decir, forma parte esencial del sentido del término 'conjunto'. Si casos como el del término 'conjunto' tiene cierta plausibilidad es que hay algún problema en la argumentación de Noonan. Es decir, esta especie de empate

entre relativistas y absolutistas que Noonan quiere defender, en lo que a criterios de identidad se refiere, es discutible.

Para el relativista no tenemos más que relaciones de equivalencia relativas a partir de las cuales se generan términos contables. Pero teniendo en cuenta que toda relación se define en un dominio, parece indudable que al especificar un dominio, tanto para la interpretación de un lenguaje formal como para los diferentes ejemplos que los relativistas plantean, hay una relación de identidad absoluta asociada al mismo. Esto es precisamente lo que el relativista niega. Parece ser que el relativista considera que el dominio no se especifica de esta manera, sino más bien a través de algún procedimiento como el de las listas de Geach. Desde el punto de vista relativista no se trata de que tengamos, por ejemplo, el dominio de los animales individuales y sobre él definamos o podamos definir diferentes relaciones de equivalencia. Si así fuera, una vez más, la identidad absoluta se presentaría. Parece ser que el dominio es algo “carente de forma” y que no sólo *ser la misma especie animal* que es una relación de equivalencia sobre ese dominio informe, sino que *ser el mismo animal individual* que también lo sería. De alguna manera estas relaciones van conformando de una determinada manera el dominio informe. Es más, cabría pensar en otras relaciones más finas que *ser el mismo animal individual* que. La solución del caso Tibbles parece ser acorde con lo ahora señalado. Tenemos un único gato, ya que para contar los gatos que hay tenemos que considerar la relación *ser el mismo gato* que. Sólo tenemos un gato, porque Tibbles-menos-un-pelo, Tibbles-menos-dos-pelos, etc., son el mismo gato que Tibbles. A efectos de cuenta sólo hay un gato, a pesar de que Tibbles-menos-un-pelo es distinta materia continua de tejido felino que Tibbles-menos-dos-pelos. Es como si la relación *ser el mismo gato* que se definiera sobre un dominio de materias continuas de tejido felino. ‘Tibbles’ en realidad sería un nombre-para; una especie de nombre *compartido* por todos los elementos de la clase de equivalencia que viene a constituir un gato individual. Quizás quepa pensar que ‘Tibbles-menos-un-pelo’ siendo el nombre de un gato (no el nombre-para un gato) pueda ser el nombre-para una masa continua de tejido felino, y así sucesivamente. Un dominio informe se va reconfigurando visto desde diferentes relaciones de equivalencia. Aquellas listas de nombres propios pretenden representar este dominio informe que no necesita de ninguna relación de identidad asociada. El análisis relativista del caso Tibbles puede ser criticado desde diferentes frentes. Hay argumentos “ontológicos” que ponen en duda el análisis geachiano del caso Tibbles. La idea básica es aquella que Wiggins popularizó en los medios filosóficos: Tibbles-menos-un-pelo no es el mismo gato que Tibbles por la sencilla razón de que no es un gato, interpretando que el referente de ‘Tibbles-menos-un-pelo’ es la masa continua de tejido felino, menos un pelo, que *constituye* Tibbles.

A lo largo de este trabajo lo que se ha tratado de poner en cuestión es que este esquema lingüístico-ontosemántico que se genera a partir del relativismo, por muy consistente que pueda ser, se ajuste a nuestra práctica ordinaria. Por supuesto, se está asumiendo que la discusión planteada en el terreno de los lenguajes formales alguna relación tiene con esa práctica ordinaria. No parece que el esquema relativista se ajuste a los hechos lingüístico-ontosemánticos situados en el nivel que los lenguajes formales y su semántica representan, ni al nivel en el que hacemos uso de términos generales y nombres propios. Noonan, al hacer una valoración del relativismo (en Noonan (1997)), nos recuerda que Dummett ha propuesto una distinción plausible entre un nivel inferior y otro superior dentro del lenguaje. Asimismo, Dummett describe brevemente como se da esa transición entre los dos niveles. Aquí aparecen unos enunciados parecidos a los de identidad. Dice 'parecidos' ya que mediante ellos no se expresan relaciones entre objetos: en ese nivel inferior no hay objetos constituidos. Noonan parece querer decirnos que esta concepción es bienvenida por parte del relativista, puesto que mostraría que la relación de identidad absoluta no tiene ningún sentido al no haber un dominio de objetos constituidos. Desde nuestro punto de vista, parece que Noonan está arrinconando un tanto al relativista. No parece que en la discusión mantenida al nivel de los lenguajes formales las concepciones dummettianas sobre los distintos niveles tengan relevancia. Quizás, si la concepción dummettiana fuera correcta, el relativista tendría su hueco. Pero esto es desviar la discusión a otro campo de batalla.

Para finalizar, un breve apunte sobre identidad relativa y relatividad ontológica. La relatividad ontológica implica una relatividad de la identidad en el sentido de que desde una teoría no hay medio de seleccionar uno de sus modelos. El signo de identidad está sujeto, pues, a la misma indeterminación que otros términos del lenguaje. Pero esta relatividad de la identidad no es el relativismo de Geach y sus seguidores. Este último ha sido el objeto de este trabajo.

BIBLIOGRAFÍA

- Alston, W.P. and Bennett, J.: "Identity and cardinality: Geach and Frege", *The Philosophical Review* (1984), XCIII(4), 553-567.
- Dummett, M.: *Frege. Philosophy of Language*. Cambridge, Mass.: Harvard University Press, 1973 (segunda edición, 1981). Cap.16: Identity.
- Dummett, M.: *The interpretation of Frege's Philosophy*. Cambridge, Mass.: Harvard University Press, 1981. Cap.11: Geach on identity.
- Dummett, M.: "Does Quantification Involve Identity?". En Lewis (1991).
- Frege, M.: *Die Grundlagen der Arithmetik*. 1884. Traducción de Ulises Moulines en Frege, G.: *Escritos Filosóficos*. Barcelona: Crítica, 1996.
- Geach, P.: *Reference and Generality*. Londres: Cornell University Press, 1962 (tercera edición, 1980).
- Geach, P.: "Identity", *Review of Metaphysics* (1967), 21(1).
- Geach, P.: "Identity: a reply", *Review of Metaphysics* (1969), 22(3).

- Geach, P.: "Ontological relativity and relative identity". En Munitz, M.K. (ed.): *Logic and Ontology*. New York: New York University Press, 1973.
- Geach, P.: "Replies". En Lewis (1991).
- Lewis, H.A. (ed.): *Peter Geach: Philosophical Encounters*. Dordrecht: Kluwer, 1991.
- Lowe, E.J.: "Objects and criteria of identity". En Wright-Hale (1997).
- Noonan, H.: "Absolute and relative identity". En Williams, S.G. and Lovibond, S. (eds.): *Essays for David Wiggins*. Oxford: Aristotelian Society Series (16), 1996.
- Noonan, H.: "Relative identity". En Wright-Hale (1997).
- Quine, W.V.O.: *Ontological Relativity and other essays*. New York: Columbia University Press, 1969.
- Wright, C. and Hale, B. (eds.): *A Companion to the Philosophy of Language*. Oxford: Blackwell, 1997.

Agustín Arrieta Urtizberea
UPV-EHU
e-mail: ylparura@sf.ehu.es





CONTRAFECTICIDAD Y DERROTABILIDAD¹

ESTANISLAO BARRY

Universidad de Buenos Aires

SANDRA LAZZER

Universidad de Buenos Aires

ABSTRACT: The main goal of this paper is to analyze the notions of *counterfactual and defeasible character* attributed to certain type of expressions. Firstly, we will put forward some of the fundamental subjects related to each of these notions. Then, we will discuss several problems that arise with a possible combination of both notions in only one conditional expression. The loss of the *strengthening antecedent law* and the validity of *conditional modus ponens* are central matters for conditional logic. Beginning with an analysis of the distinctive formal properties presented by the different conditional logic systems, we will give place to the discussion about the properties that a defeasible counterfactual conditional should have.

KEYWORDS: Conditional Logic, counterfactual conditionals, defeasibility.

I

En situaciones cotidianas en las cuales tratamos de pensar razonablemente acerca de un problema práctico usualmente estamos obligados a elegir entre diversas alternativas. Pensar razonablemente acerca de un problema significa, entre otras muchas cosas, explorar en detalle las consecuencias de cada una de las alternativas a considerar. La representación de este tipo de proceso deliberativo y/o evaluativo se realiza en el ámbito del *razonamiento del sentido común*. Tanto nociones no deductivas de inferencia, como ciertos tipos especiales de operadores condicionales, son instrumentos formales que se han usado en este ámbito. Cuando para caracterizar en términos lógicos cierto tipo de expresiones lingüísticas de la forma “si...entonces” usadas en el lenguaje natural se introducen un operador condicional, hay por lo menos dos posibles situaciones que se intentan

¹ Este trabajo se hizo en el ámbito del proyecto de investigación UBACYT FI-105 “Nociones de Consecuencia No Estándares y Lógicas Condicionales”. Los autores agradecen a David Makinson por sus importantes comentarios.

reflejar. Por un lado en ciertas circunstancias podríamos querer determinar cuál habría sido el caso si cierta posible situación o hecho hubiera acontecido, situación ésta que puede expresarse por medio de un enunciado de la forma '*si se hubiera dado el caso, se habría dado el caso*'.

Por otro lado, hay circunstancias en las cuales la consideración de presuposiciones ligadas a contexto, o la incorporación de nueva información relevante podría modificar aquello que estamos dispuestos a aceptar como una consecuencia razonable. Enunciados como '*si en principio se da tal o cual cosa o si bajo tales o cuales condiciones se da tal o cual cosa, entonces sucede o se dará tal y cual otra*' son enunciados condicionales presentes en nuestros procesos evaluativos. Ahora bien, mientras que en el primer tipo de enunciados se lo suele identificar un operador condicional de tipo *contrafáctico*, el segundo se lo asocia con la presencia de un condicional *derrotable*². Pero ¿pueden estas dos nociones combinarse en una misma afirmación condicional? ¿Y qué tipo de enunciado condicional es aquel que podríamos denominar *contrafáctico derrotable*?

En este trabajo nos proponemos discutir alguna de las dificultades intuitivas y formales para una posible compatibilización de los conceptos de *contrafácticidad* y *derrotabilidad*. Para ello presentaremos, en primer lugar, alguna de las cuestiones ligadas a las lógicas que dan cuenta de estas dos nociones. En segundo lugar trataremos de esclarecer algunas confusiones inherentes a la compatibilidad e incompatibilidad de ciertas propiedades que caracterizan a cada uno de estos tipos de condicionales, para finalmente esbozar cuáles son los problemas a resolver, si de formular una lógica para condicionales *contrafácticos derrotables* se trata.

II

Entre los usos más frecuentes que le damos a cierto tipo de expresiones condicionales que denominadas *contrafácticos* están las siguientes: representar un razonamiento hipotético y justificar o explicar una acción (o una inacción); dar cuenta de varios tipos de necesidad y de la noción de causalidad; representar estados de creencias y de conocimiento; dar cuenta de la noción '*acción libre*' que suele formularse en términos de estos condicionales: por último analizar la noción de disposición y representar la idea de ley de la naturaleza, no como un enunciado meramente verdadero, sino como una sustentado por un condicional de este tipo. Pero, a pesar de la amplia gama de aplicación que la noción de *contrafácticidad* parece tener, no es sin embargo fácil establecer una clara identificación de estos condicionales a nivel del lenguaje natural. *contrafácticos* a veces se contraponen a condicional indicativo, pero a la vez y para conservar la caracterización

² Traducimos aquí, siguiendo a C. Alchourrón *defeasible* por *derrotable*. Traducciones alternativas son *rebatible* y *revocable*.

gramatical como criterio clasificatorio, contrafáctico se asimila a la noción de condicional subjuntivo. Consideraciones de distinto tipo suelen mezclarse en estas distinciones. *Contrafactividad* denota una propiedad de la oración condicional, lo que podríamos denominar el *carácter contrafactico* de la oración, al ser afirmada en un contexto apropiado. La distinción *subjuntivo* e *indicativo* se refieren a la forma gramatical de dicha afirmación.

A veces resulta sorprendente las dificultades existentes para caracterizar las diferencias entre los dos condicionales que solemos llamar indicativos y contrafácticos. No hay siempre un acuerdo en que nivel (sintáctico, semántico o pragmático) se da la discusión. En muchos casos estos intentos clasificatorios a nivel del lenguaje natural, parece no tener un interés intrínseco, en el sentido de ser un problema a resolver por desde un punto de vista lingüístico o de la filosofía del lenguaje. Parece ser una cuestión extrínseca a estas disciplinas, que tal vez los lógicos han impuesto, en la medida en que la 'justificación pragmática de una lógica por el ámbito del lenguaje corriente que logra reconstruir'³ afectaría la adecuación entre los sistemas formales y las expresiones del lenguaje natural que con ellos se intenta reconstruir. Cualquiera sea la respuesta que se vaya a dar al problema clasificatorio, una vez estipulado, aunque más no sea por medio de un o un conjunto de ejemplos que se tomen como paradigmáticos, qué se considera un contrafáctico, la cuestión a resolver será la de saber qué estructura formal (conectivo u operador condicional) representa este condicional, y seguidamente determinar qué propiedades son las que lo caracterizan.

'Si *A* entonces *B*' es la estructura típica de una oración condicional. El tratamiento clásico traduce las oraciones de esta forma por un condicional material. Un condicional material, representado por $A \rightarrow B$ es una fórmula cuyo valor de verdad se determina en función del valor de verdad de sus partes componentes y dada la definición semántica del mismo, resulta ser verdadero siempre que su antecedente *A* sea falso. Tal vez algunos condicionales (o todos) formulados en indicativo puedan representarse por el condicional material, pero lo que parece claro es que aquellos formulados en subjuntivo que involucran algún tipo de carácter contrafáctico no pueden representarse de esta manera con un condicional material. Esto hace que, la verdad falsedad de un condicional contrafáctico no sea función de verdad de los valores de verdad de su antecedente y consecuente⁴ Para resolver este y

³ Alchourrón, C.: 'Concepciones de la lógica', en *Enciclopedia Iberoamericana de Filosofía*, volumen 7, pág 45.

⁴ Aunque el antecedente del siguiente condicional sea falso no por ello estaríamos dispuestos a aceptar la verdad del siguiente condicional 1) Si Cervantes no escribió el *Quijote*, entonces algún otro lo hizo

Pero si cambiamos el modo verbal en que la oración está formulada, pasando del indicativo al subjuntivo, obteniendo así, 2) Si Cervantes no hubiera escrito el *Quijote*, entonces algún otro

otros problemas ligados al tratamiento formal de expresiones condicionales, hacia fines de los 60' se comenzaron a formular sistemas de lógica condicional. Estos sistemas que caracterizan tanto la noción de prueba como la semántica para una lógica que contiene ahora un nuevo operador condicional, usualmente simbolizado como ' $>$ '. Con este operador se pretende formalizar de algunos usos del lenguaje natural de expresiones condicional. Los sistemas de lógica condicional propuestos por autores como Stalnaker, Lewis y Pollock, tienen estas características.

Una lógica condicional (contrafáctica) es un cierto tipo de sistema intensional que puede ser visto como una extensión de la lógica proposicional clásica. Contiene el conjunto de todas las tautologías y está cerrada bajo *modus ponens*: $((A \rightarrow B) \wedge A) \rightarrow B$. El nuevo operador intensional ' $>$ ' se caracteriza a través de un conjunto de axiomas y reglas de inferencia. Una característica común a todas las presentaciones de sistemas de lógica para condicionales especiales es que en ninguno ellos se valida la *ley del refuerzo del antecedente* para el operador condicional. $(A > B) \rightarrow ((A \wedge C) > B)$. Pero es una característica común a todos los sistemas de condicionales contrafácticos la presencia del *modus ponens condicional*: $(A > B) \rightarrow (A \rightarrow B)$.⁵ Este tipo de sistemas recibió también un tratamiento semántico en términos de una semántica de mundos posibles. Estas son semánticas kripkeana a las que se le agrega una función ordenadora entre mundos. Las dos versiones más difundidas de este tipo de semántica fueron son la que utiliza una *función de selección* (Stalnaker) y la que introduce un *sistema de esferas* (Lewis). Lo que la función selección o los sistemas de esferas representarán es la idea de *mínima diferencia (o máxima similaridad) respecto de un mundo determinado en el que la evaluación tiene lugar*. Aunque formalmente no sea un requisito considerar como actual (destacado) el mundo donde se realiza la evaluación, tanto en los modelos de Stalnaker como los de Lewis esta presente esta noción. Para Stalnaker las condiciones de verdad de un condicional del tipo $A > B$ esta fijadas por el mundo en que A es verdadero que difiere mínimamente del mundo actual. Para Lewis el valor de verdad de un condicional contrafáctico no depende del valor de verdad que adopte el consecuente B en todos los mundos de un cierto tipo en el que A vale, sino el valor de verdad de B en los mundos (en todos ellos) más parecidos al actual en los que vales A . La teoría semántica ofrece entonces una precisión formal de la noción de similaridad comparativa a un mundo dado.

Un elemento importante es que tanto las modelos semánticos de Lewis como los de Stalnaker tratan al condicional contrafáctico como *estricto en por*

lo habría hecho resulta por lo menos improbable que aceptemos como verdadera a esta segunda oración.

⁵ Algunas presentaciones de sistemas para condicionales contrafácticos son el sistemas el VC de Lewis, el sistema C2 de Stalnaker y el sistema SS de Pollock.

lo menos dos sentidos. Por un lado se trata de un tipo especial de estrictez que se diferencia de lo que en la lógica modal clásica es la implicación estricta. Según Lewis lo estricto que es un contrafáctico varía al evaluar el valor de verdad del mismo en un secuencia de situaciones posibles, lo que vuelve inaceptable la posibilidad de reforzar el antecedente. Pero además, un contrafáctico es estricto en el sentido de que no admite ninguna excepción. Un condicional $A > B$ es verdadero en un mundo w si B es verdadero en *todos los mundos*, sin excepción, entre los mundos más similares en los que A es verdadero, de lo que resulta la validación inmediata del *modus ponens condicional*.

Entre los distintos sistemas propuestos por Lewis está el que este autor consideró apto para representar la noción de obligación condicional conocido como sistema V . Entre las diferencia sintácticas con el sistema que representa en Lewis la noción de contrafactividad, esto es el VC , se encuentra el que V no incluye entre sus tesis al *modus ponens condicional*. En términos semánticos esto se refleja por el relajamiento de ciertas condiciones como es el centramiento que son indispensables para la validación del *modus ponens condicional*. Es preciso señalar que este tipo de condicionales serán los que se asocian con la noción de derrotabilidad. Esta noción sin embargo se aplica tanto a un operador condicional, obviamente en el lenguaje objeto, como a una noción de inferencia (meta lingüística). Cuando se habla de condicionales derrotables, se suele asociar este concepto a un operador que no tiene *modus ponens*, en cambio cuando se aplica la noción para caracterizar una inferencia se la asocia a la pérdida de una propiedad que la noción de inferencia deductiva tiene, esto es la monotonía. La no-monotonía también fue asociada con lo que denominamos ley del refuerzo del antecedente. ¿Cómo se combinan todas estas características? ¿Cuáles resultan compatibles y cuáles incompatibles? Para poder dar una respuesta a esto necesitamos antes introducir algunos rasgos característicos de lo que vamos a entender por derrotabilidad.

III

Derrotable y *derrotabilidad* son expresiones que originalmente estuvieron ligadas al ámbito de la ética y de la filosofía del derecho. En estos ámbitos las acciones suelen ser evaluadas como correctas o incorrectas, justificadas o injustificadas, obligatorias o prohibidas; los individuos mismos pueden ser evaluados como buenos o malos, morales o inmorales. La ética puede ser vista como la rama de la filosofía que estudia este tipo de procesos evaluativos. En la filosofía del derecho la noción de *prima facie* se utiliza para caracterizar cierto tipo de obligaciones que bajo ciertas circunstancias pueden ser abandonadas. Suelen hacerse también evaluaciones de estados de creencias y de otros procesos cognitivos: las creencias pueden estar justificadas o no. También respecto de los individuos podemos decir que son (o actúan)

razonablemente o no. La epistemología puede ser entonces descripta como una rama de la filosofía que estudia los principios generales que gobiernan nuestras evaluaciones epistémicas. En el ámbito de la teoría del conocimiento, y como una posible respuesta al problema planteado por E. Gettier, se introdujo el concepto de derrotabilidad para caracterizar una nueva forma de entender el proceso de justificación epistémica. Por otro lado, los conceptos de derrotable y derrotabilidad llegan al ámbito de la lógica tanto desde los desarrollos teóricos de la inteligencia artificial, como de las investigaciones en el área denominada lógica filosófica. En el primer caso están ligados a los estudios lógicos formales del razonamiento del sentido común y en el segundo caso pretendieron dar cuenta en términos de un sistema de lógica del concepto obligación *prima facie*. También en otros ámbitos, como el de filosofía de la ciencia, hay una noción ligada a este concepto (en el sentido que podría ser entendida en términos de derrotabilidad) como lo es la noción de condición o cláusula *ceteris paribus*. Otros conceptos importantes en la literatura filosófica como lo son el de *condición de posibilidad*, *contenido implícito* y *presuposición* también guardan una estrecha relación con la noción de derrotabilidad.

Ahora bien, esta lista, por cierto no exhaustiva, de ámbitos donde se ha aplicado el concepto de derrotabilidad nos lleva a considerar la cuestión más general de qué es la derrotabilidad.⁶ Una primera distinción consistirá en diferenciar los casos en que el concepto de derrotabilidad se aplica a la conclusión de un argumento, de los casos en los que se aplica a la noción misma de inferencia y por ende caracteriza un tipo de argumento. Usualmente se ha aplicado el concepto de derrotabilidad a la conclusión de ciertos argumentos en el contexto de la ética y la filosofía del derecho. En un proceso argumentativo mediante el cual queremos justificar la adopción de una cierta norma o principio podemos concluir que el tipo de obligación involucrada en ese principio es una obligación *prima facie*.⁷ En cambio, en el ámbito de la lógica el término fue aplicado para caracterizar a cierto tipo de razonamientos o inferencias. En términos muy laxos e introductorios, un razonamiento puede ser considerado derrotable si resulta ser nulo o debe ser abandonado en virtud de ciertas circunstancias que no están en general especificadas en las premisas.⁸ Estas circunstancias suele no estar al mismo nivel que las premisas,

⁶ En lo que sigue usaremos la presentación del concepto ofrecida por D. Makinson [1993], introduciendo sin embargo algunas modificaciones

⁷ La idea de obligación *prima facie* fue tratada formalmente en la lógica deóntica bajo la forma de un operador diádico de obligación, i.e. de un operador de obligación condicional, cuya forma típica es '*si A, entonces debe ser el caso que B*'

⁸ Nótese que bajo esta definición podrían también quedar comprendidos procesos inferenciales como el de algunas teorías de la inducción, la teoría peirceana de abducción y el concepto de inferencia a la mejor explicación

pero están asociadas tanto a la inferencia como al contexto donde el razonamiento se formula.

D. Makinson en 'Five Faces of Minimality', distingue por lo menos tres de situaciones a partir de las cuales podemos considerar que un razonamiento o una inferencia es derrotable. Hay un primer caso en el que la derrotabilidad de una inferencia consiste en la capacidad que tienen otras inferencias para derrotarla. Una inferencia será considerada derrotable cuando no pueda seguir siendo aceptada sí, al enfrentarse a un contrargumento, éste nos brinda evidencia contraria a lo que el argumento pretendía concluir. Esta noción está presente en los sistemas *razonamientos derrotables* (defeasible reasoning) propuestos por D. Nute, J. Pollock y otros, todos ellos inspirados en el tratamiento filosófico de esta noción sugerido por R. Chisholm. La noción básica a caracterizar es la de cuál será el elemento que se considere *derrotador* (defeater). En un segundo caso tenemos una noción de derrotabilidad involucrada en sistemas como las *lógicas default*, las de la negación por falla de Porlog y las lógicas autoepistémicas, propuestas por R. Reiter, R. Moore y otros. En este segundo caso la derrotabilidad está relacionada con una pérdida de la consistencia por el añadido de ciertos elementos del contexto que invalidan la conclusión. Todos estos sistemas tratan con la generación de una noción de inferencia derrotable, que tiene por objetivo impedir que la conclusión se vuelva inconsistente frente al agregado de nueva información a una *base de conocimiento*, y se basan en una tipo prueba de consistencia que incorpora a esta base todas las fórmulas no deductivas que sean consistentes con ella. El tipo de reglas de inferencia que estas lógicas proponen permiten saltar a conclusiones no establecidas deductivamente a partir de las premisas. Pero hay un tercer caso, que es centro de atención en el artículo de Makinson, donde la derrotabilidad se caracteriza como una falta de *normalidad* en las situaciones a partir de las cuales estamos dispuestos a considerar verdaderas a las premisas de un argumento. Para aclarar esto debemos introducir algunas precisiones.

El estudio de la noción de inferencia derrotable desde fines de los 70' está ligado al desarrollo de las llamadas *lógicas no-monótonas*. Las teorías formales que estudian los procesos inferenciales no-monótonos (entre las cuales se deben incluir las mencionadas anteriormente) difieren entre sí no sólo en el sentido que le dan al concepto de derrotabilidad, si no también si no en el nivel de generalidad que eligen para formular la teoría. La noción de inferencia derrotable es caracterizada en estas lógicas en términos negativos: como la falla de la *propiedad de monotonía*. Aunque la derrotabilidad y consecuentemente lo no-monotonía haya sido entendida como una propiedad atribuible a ciertos casos de argumentos específicos, pronto cobró interés el estudio general de la relación de inferencia derrotable, como un tipo especial de relación de inferencia. D. Gabbay fue el primero en dar cuenta de cuáles eran las propiedades formales que podrían caracterizar una relación de

inferencia no-monótona, donde el punto de partida para el análisis fue la consideración del caso de la relación estándar deductiva propuesta por Tarski. La nueva relación de inferencia se caracteriza como una relación de *inferencia cumulativa* (cumulative inferece) la que introduce un tipo especial de monotonía que D. Makinson bautizó como '*cautelosa*' (cautious monotony). Este tipo de inferencia pudo ser presentado en términos de cierto tipo especial de modelos denominados *modelos preferenciales*. Esta es una semántica de tipo modal donde la noción básica que establece una relación de orden entre los mundos es la noción de normalidad. Entonces los que caracteriza a una inferencia derrotable es establecer cuál es la mínima distancia que nos separa de una completa o ideal normalidad. *Una proposición B podrá inferirse de otra proposición A, si B es verdadera en todo mundo que es mínimamente anormal de aquellos que satisfacen A.*

Este tipo de semánticas también puede utilizarse para dar un modelo a un operador condicional del lenguaje objeto. Si consideramos la proposición condicional $A > B$ podemos transformar la regla que caracteriza semánticamente a la relación de inferencia derrotable en una cláusula de valuación para las fórmulas de tipo $A > B$. Esto no puede resultar sorprendente si se tienen en cuenta ciertas similitudes sintácticas entre algunos sistemas de lógicas condicionales y de lógicas no-monótonas. Es evidente, por ejemplo, que la inferencia clásica representada por la ley de refuerzo del antecedente, la transitividad (o silogismo hipotético) y la contraposición que fallan en las lógicas condicionales, también fallan a un nivel meteórico respecto de una noción alternativa de consecuencia lógica como es la no-monótona. Esto a hecho que sea común encontrar en la literatura la idea de que la falla del refuerzo del antecedente para un conectivo condicional es la contraparte de la propiedad de no-monotonía de la noción de consecuencia de este tipo de lógicas. Sin embargo, algunas lógicas condicionales, como la de los condicionales contrafácticos, valida ciertas propiedades formales que se presentan como indeseables, si de dar cuenta de la noción de default y/o de la noción de *prima facie* se trata. Específicamente lo que hemos denominado *modus pones condicional* es una propiedad que no está presente en ninguna lógica que de condicionales derrotables, sea con intereses centrado en la noción de *default* (Delgrande)⁹ que viene de la AI o sea centrado en la noción de *prima facie* que viene de la lógica deóntica (Hansson)¹⁰. En algún sentido la noción de derrotabilidad asociada al operador condicional impone la falla del *modus pones condicional*. Pero entonces ¿qué sentido podría tener la combinación del carácter contrafáctico y el carácter derrotable en un mismo

⁹ Delgrande, J.: "A Frierst Order Conditional Logic for Prototypical Properties" *Artificial Intelligence*, 33 (1987)

¹⁰ Hansson, B.: "An Analysis of some Deontic Logics" en R. Hilpinen (ed) *Deontic Logic* D. Reidel, Dordrecht. (1971)

condicional? En el último apartado nos ocuparemos de aclarar algunos errores en la presentación de esta cuestión, para mostrar finalmente en qué sentido es viable una combinación de estas dos nociones y como debería encararse el problema de formular un sistema de lógica que de cuenta de un operador condicional con estas características.

IV

¿Hay contrafácticos derrotables? De lo que acabamos de presentar podrían seguirse dos posibles respuestas: a) Todos los contrafácticos son derrotables; b) Ningún contrafáctico es derrotable. Estas dos respuestas, por cierto erróneas, se basan en una suerte de confusa asimilación o identificación de conceptos como ‘no-monotonía’, ‘derrotabilidad’, ‘ley del refuerzo del antecedente’ y ‘modus ponens condicional’ que intentaremos aclarar.

En primer lugar, tal como vimos en el apartado anterior, la formulación de sistemas de lógica no-monótonas parece ser una de las formas con que se ha dado cuenta de la noción inferencia derrotable, pero no es la única, con lo cual la identificación sin más aclaraciones de los dos conceptos lleva a malos entendidos. En segundo lugar, aunque la pérdida de la ley de refuerzo del antecedente sea una forma de reflejar en el lenguaje objeto la no-monotonía, esto no significa que sea ésta la única forma que adquiere la no-monotonía, tal como lo muestra claramente Makinson en ‘Five Faces of Minimality’, ni que los sistemas de lógica condicional cuyos operadores no validen ésta ley deban ser considerados *per se* no-monótonos. Las lógicas condicionales en tanto extensiones de la lógica clásica proposicional están, hablando en un sentido metafórico, montadas sobre una noción de consecuencia clásica y deductiva (Chellas¹¹) y por ello sería en este sentido erróneo decir que son no-monótonas. Ahora, esto no significa que, no exista la posibilidad de *representar* la noción de no-monotonía a nivel del lenguaje objeto mediante un operador condicional (deductivo). En tercer lugar, cuando se dice que *todos los condicionales contrafácticos son derrotables* se puede cometer el siguiente error: dado que los contrafácticos carecen de refuerzo del antecedente, éste caracteriza la no-monotonía y ésta última identifica a la propiedad de derrotabilidad, entonces todos los contrafácticos son derrotables. Obviamente si se hace esta identificación los contrafácticos resultan ser derrotables en un sentido trivial. Pero además esto ocasiona problemas a la hora de proponer un operador condicional al que llamamos ‘derrotable’ *per se*, es decir un operador que refleje lo que intuitivamente está presente en la idea de un *default* y un enunciado *prima facie*. En cuarto lugar entonces tendríamos el problema de que estos sistemas de lógica condicional donde se ha querido representar estas nociones no validan, a diferencia de los

¹¹ Chellas, B.: "Basic Conditional Logic" en *Jouranal of Philosophical Logic*, 4 (1975) pp 133-153.

contrafácticos, el *modus ponens* condicional. En esto se basa la afirmación también errónea de que *ningún contrafáctico es derrotable*. Aquí tal vez más que de un error se trate de una identificación abusiva, según la cual un condicional derrotable se caracteriza esencialmente como aquel que no valida el *modus ponens* condicional. El malentendido puede aumentar si se destaca además que la pérdida del *modus ponens* condicional implica la pérdida de un tipo especial de ley del refuerzo del antecedente, la que C. Alchourrón llamó *ley del refuerzo material del antecedente*:

$(A > B) \rightarrow ((A \wedge C) \rightarrow B)$. Makinson mostró que ésta se corresponde con una forma de monotonía que denominó *pseudo monotonía*. La ley de refuerzo del antecedente presentada por Lewis, i.e. $(A > B) \rightarrow ((A \wedge C) > B)$ (que ningún condicional especial valida) y el *modus ponens* condicional son independientes, en el sentido de que la pérdida (ni la validación) de una implica la validación de la otra. Pero si por definición, como pasa en el sistema DFT propuesto por Alchourrón, ‘derrotable’ (por definición) es un condicional que no valida el *modus ponens* condicional¹², entonces *ningún contrafáctico es derrotable*. ¿Pero entonces derrotabilidad y contrafácticidad son dos nociones que no pueden combinarse nunca? Identificar derrotabilidad con pérdida del *modus ponens* condicional es abusivo en el sentido de que deja fuera ciertos ejemplos intuitivos, como el propuesto por Makinson donde las dos nociones parecen combinarse. Por último, presentaremos este ejemplo, esbozando sólo en términos introductorios cuáles son las características formales que se esperaría de una lógica que represente este tipo de operador.

V

‘Si Mozart hubiera vivido hasta 1800, habría escrito (definitivamente) algunas sinfonías más’. Este es el enunciado con que Makinson presenta la noción de contrafáctico derrotables. Pero ¿por qué le atribuye el sentido de ser un *contrafáctico derrotable*?. Imaginemos la siguiente situación: se descubre que Mozart con el fin de escapar de sus acreedores falsificó el dato de su fallecimiento, fijándolo en 1791, viajando después a la Patagonia donde se establece como un obsesivo pero poco exitoso buscador de oro, si componer una sola línea más hasta el momento de su muerte en 1810. Parece entonces que no estamos más en una situación (Mozart continuando su vida en Salzburgo en circunstancias normales) como la que podríamos haber supuesto de no contar con esta información adicional.

Las llamadas semánticas de la normalidad también se basan en la idea de introducir una función ordenadora entre mundos, pero a diferencia de las de la similaridad comparativa usadas para los contrafácticos, ordenan los

¹² El condicional derrotable del sistema DFT tampoco valida la ley de refuerzo del antecedente de Lewis, pero es la pérdida del *modus ponens* condicional lo que lo distingue y diferencia, un condicional contrafáctico

mundos de acuerdo con una distancia a un mundo más ideal o completamente normal. En "Five Faces of Minimality" Makinson analiza distintas versiones de lo que en la literatura se ha dado en llamar *semánticas de la normalidad*. Presenta primero la noción de una semántica de la *normalidad absoluta* pero para caracterizar una inferencia o consecuencia derrotable y no un conectivo del lenguaje objeto. La "normalidad" se entiende en un sentido absoluto porque lo que cuanta como normal o típico es independiente del mundo en cuestión. Se puede extender este tipo de semánticas a lenguajes donde ocurran proposiciones del tipo $A > B$, dando así modelos para el condicional especial. Makinson presenta además una variante dentro de estas semánticas, como son las que denomina semánticas de la *normalidad variables*, donde en términos muy generales, la diferencia con las anteriores está en que aquello que se considera "normal" varía de acuerdo con mundo desde donde la situación es vista¹³. Ahora bien, tal como dijimos en el apartado II, lo que la función selección o los sistemas de esferas (en los sistemas de Stalnaker y Lewis) representan es la idea de *mínima diferencia (o máxima similaridad) respecto de un mundo determinado en el que la evaluación tiene lugar*. Tenemos entonces, en términos intuitivos, dos tipos de semánticas, de la *normalidad* y de la *similaridad*, cuya diferencia básica está en la noción intuitiva sobre la que se construye la función ordenadora, mayor cercanía (o menor distancia) en el primer caso, y mayor normalidad en el segundo.

Si volvemos sobre la noción de contrafáctico derrotable que aquí nos compete, se presenta ahora como posibilidad, y en esto consiste la propuesta de Makinson, entender a esta noción en función de un modelo semántico que tiene la particularidad de combinar estos dos tipos de ordenes. Entonces $A > B$ es un *contrafáctico derrotable* si i) es verdadero en un mundo w que es el más cercano entre los más normales que satisfacen A , o ii) es verdadero en todo mundo w tal que es el más normal entre los más cercanos que satisfacen A . En ninguno de los dos casos se valida el *modus ponens condicional*. Si bien es cierto que presentar el problema de qué es un contrafáctico derrotable por este camino evita mucho de los problemas y malos entendidos que reseñamos antes, deja abierto aún ciertas dificultades formales e intuitivas si quisiéramos construir un sistema de lógica para este tipo de condicionales. Para terminar esbozaremos de manera muy general e introductoria sólo dos de estas cuestiones.

En primer lugar no disponemos de un criterio claro para distinguir en el ámbito de los usos de las expresiones contrafácticas en el lenguaje natural, cuando un hablante formula un contrafáctico con un sentido derrotable y cuando no. ¿Hay contrafácticos que sean "inderrotables"? Algunos han visto

¹³ Con este tipo de semánticas llamadas de la normalidad variable se puede introduciendo una restricción especial dar cuenta de los sistemas clásicos para condicionales contrafácticos, como por ejemplo el VC de Lewis

en expresiones como "si hubiera tenido tres lados habría sido un triángulo" un ejemplo de este último tipo de condicionales. Pero en principio es dudoso que estos condicionales puedan ser contrafácticos en algún sentido interesante. Esta cuestión no afectaría directamente la formulación de un sistema de lógica pero si trajese dificultades a la hora de evaluar la supuesta adecuación (pragmática) con las expresiones del lenguaje natural que se pretende (si es esto lo que se pretende hacer) reconstruir.

En segundo lugar podríamos preguntarnos ¿qué propiedades formales validan los contrafácticos derrotables que los distinguen del resto de lo derrotables? Si se tiene en cuenta la el peso que tiene la presencia del modus ponens condicional en los sistemas clásicos de contrafácticos y la incompatibilidad manifiestas entre la satisfacción de esta regla y la noción de derrotabilidad tal vez podría ser plausible esperar que algo al que denominemos contrafáctico derrotable no valide ya el modus ponens condicional clásico pero si una forma que podríamos denominar débil del mismo, esto es $(A \wedge (A > B)) > B$.

Esta fórmula tiene la particularidad de ser de segundo orden en el sentido que dentro del alcance del operador $>$ ocurre otro operador $>$. En la literatura se han propuesto algunos sistemas cuyo operador valida esta fórmula ¿son ellos los que representan la noción de contrafáctico derrotable?.

Tal vez la "derrotabilidad" pueda ser vista como una propiedad general que puede en ciertas circunstancias predicarse de cualquier tipo de operador condicional, no sólo de los contrafácticos. Pero aun así queda abierta la posibilidad de una combinación con la noción de condicional contrafáctico y por lo tanto surge el problema no resuelto de cuál es el sistema de lógica que da cuenta de estos condicionales.

BIBLIOGRAFÍA

- Alchourrón, C. E. (1995): "Concepciones de la lógica", en *EIAF*, Tomo7, Lógica, Ed. Trotta.
- Alchourrón, C.E. (1995): "Defeasible Logic: Demarcation and Affinities", en *Conditionals: from Philosophy to Computer Science*, edited by G. Crocco, L. Fariñas del Cerro and A. Herzing, Clarendon Press, Oxford.
- Boutilier, C (1992) *Conditional Logics for Default Reasoning and Belief Revision*, Technical Report 92-1 The University of British Columbia, Department of Computer Science,
- Delgrande, J.P. (1987), *A First order Conditional Logic for Prototypical Properties*, Artificial Intelligence, 33 (1).
- Kraus, S., Lehmann, D & Magidor, M. (1990): "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", *Artificial Intelligence* 44: 167-201.
- Lewis, D. [1973]: *Counterfactuals*, Cambridge, Mass., Harvard University Press.
- Makinson, D. (1989): "General Theory of Cumulative Inference", en *Second Int. Workshop in Nonmonotonic Reasoning*, Springer Verlag.
- Makinson, D. (1993): "Five faces of Minimality", *Studia Logica* 52:339-379.
- Nute, D. [1984]: "Conditional Logic", en Gabbay, D and Guenther, F., *Handbook of Philosophical Logic*, vol 2: Extension of Classical Logic, Dordrecht, Reidel, pp 387-439.
- Shoham, Y. (1987): A Semantical Approach to Nonmonotonic logics, en *Proceedings Logics in Computer Science*, Ithaca, New York.

Stalnaker, R. [1968]: "A Theory of Conditionals", en N. Rescher [ed], *Studies in Logical Theory*, Oxford, Blackell, pp 98-112.

Estanislao Barry
Universidad de Buenos Aires
e-mail: rbarry@impsat1.com.ar

Sandra Lazzer
Universidad de Buenos Aires
e-mail: slazzer@filo.uba.ar





COMBINACIONES DE LÓGICA TEMPORAL Y EPISTÉMICA*

JOSÉ RAFAEL HERRERA GONZÁLEZ
Universidad de La Laguna

MARGARITA VÁZQUEZ CAMPOS
Universidad de La Laguna

ABSTRACT: The main focus of this paper is to present some logical systems that combine temporal and epistemic operators. In order to achieve this goal, we give an overview of methods and systems that have been proposed to combine a temporal and an epistemic dimension in a logic. Finally, we consider a special system and explain its features.

There is an increasing interest in combining logics. It has also been shown how to add a temporal dimension to a logic system. The special case of having a two-dimensional temporal epistemic logic seems to be philosophically very interesting.

KEYWORDS: Temporal logic, epistemic logic, combinig logics, multidimensional logics, knowledge, belief, time.

1. Introducción

Existe un interés creciente en la combinación de distintas lógicas¹. En esta línea, se han estudiado distintas combinaciones de sistemas de lógica de una dimensión para crear sistemas bidimensionales. Así, Finger y Gabbay² analizan cómo combinar dos lógicas temporales para generar sistemas en los que coexisten dos “puntos de vista” temporales distintos. Estos autores se han centrado igualmente en el problema de la transferencia de propiedades (consistencia, completud y decidibilidad) desde los sistemas lógicos de partida al sistema combinado resultante, mostrando que hay una relación clara entre la transferencia de tales propiedades y la capacidad expresiva de la lógica resultante de la combinación.

* Este trabajo ha sido financiado con los proyectos PI1998/047 y PI1999/092 de la Consejería de Educación del Gobierno Autónomo de Canarias.

¹ Un claro ejemplo de esto es el volumen 37, número 2, primavera de 1996, del *Notre Dame Journal of Formal Logic*, dedicado al tema de “Combining Logics”.

² En Finger y Gabbay (1996).

Nuestro interés aquí se centra en la combinación de una lógica temporal y una lógica epistémica, ambas de una sola dimensión, para dar lugar a una lógica temporal epistémica de dos dimensiones, que puede ser muy interesante desde un punto de vista filosófico. Así, los resultados obtenidos con este tipo de desarrollos lógicos pueden resultar útiles para la clarificación conceptual de los distintos tipos de conocimientos y creencias que un agente tiene acerca del futuro, en función, por ejemplo, de que se trate de creencias más o menos centrales para el agente y puedan ser revisadas o abandonadas por éste con mayor o menor dificultad. También podemos valernos de estos desarrollos para dar cuenta de cómo van cambiando a lo largo del tiempo el conjunto de conocimientos y creencias de un agente o de un grupo de agentes.

La primera decisión clave a tomar en este ámbito es si queremos añadir una dimensión epistémica a un sistema temporal o una dimensión temporal a uno epistémico. En el primer caso, seremos capaces de expresar hechos epistémicos acerca de sucesos temporales (por ejemplo, “Juan cree que mañana verá a su padre”); en el segundo, expresamos hechos temporales sobre las creencias y el conocimiento (por ejemplo, “En el pasado, creía en la existencia de Dios”).

Un segundo tipo de decisión importante a tomar es la que tiene que ver con la ontología temporal que mejor se ajuste a nuestros propósitos (es decir, una concepción lineal del tiempo, o ramificada hacia el futuro, etc.)³, así como determinar qué lógica epistémica es la más adecuada (una lógica del conocimiento, de la creencia, del conocimiento y la creencia comunes, ..., o una lógica que combine varias de estas nociones)⁴. Dependiendo de estos distintos factores, vamos a tener diferentes sistemas temporal-epistémicos.

En nuestro caso, comenzaremos con un sistema epistémico de tiempo lineal mínimo e iremos añadiendo otras propiedades. Esto nos permitirá analizar algunos de los desarrollos más relevantes que se han formulado en la combinación de lógica temporal y epistémica, así como presentar nuestras propias propuestas y señalar algunos de los puntos claves en los que habrá de profundizar la investigación futura del tema.

2. Lógica del tiempo lineal

El lenguaje para la lógica temporal L_{LT} se obtiene de añadir a un lenguaje para la lógica proposicional clásica, los operadores monarios G (“*será siempre en el futuro que*”), H (“*ha sido siempre en el pasado que*”), F (“*será, al menos una vez, que*”) y P (“*fue, al menos una vez, que*”). Así tenemos que si A es una fórmula bien formada, también lo son GA , HA , FA y

³ A este respecto puede consultarse, por ejemplo, Gabbay, Hodkinson y Reynolds (1994).

⁴ Ver, por ejemplo, Meyer y van der Hoek (1995).

PA. Se cumplen, además, las siguientes definiciones metalingüísticas:
 $FA =_df \neg G \neg A$ y $PA =_df \neg H \neg A$.

En el sistema axiomático de LT, las reglas de deducción son Modus Ponens, Sustitución y Generalización de H y de G.

Los axiomas de un sistema de tiempo lineal pueden ser los siguientes:

Ax0. Todas las tautologías de la lógica clásica de proposiciones

Ax1. $G(p \rightarrow q) \rightarrow (Gp \rightarrow Gq)$

Ax2. $H(p \rightarrow q) \rightarrow (Hp \rightarrow Hq)$

Ax3. $p \rightarrow HFp$

Ax4. $p \rightarrow Gpp$

Ax5. $FFp \rightarrow Fp$ (transitividad)

Ax6. $PPp \rightarrow Pp$ (transitividad)

Ax7. $FFp \rightarrow (Pp \vee p \vee Fp)$ (linealidad hacia adelante)

Ax7'. $(Fp \wedge Fq) \rightarrow (F(p \wedge q) \vee F(p \wedge Fq) \vee F(Fp \wedge q))$

Ax8. $FFp \rightarrow (Pp \vee p \vee Fp)$ (linealidad hacia atrás)

Ax8'. $(Pp \wedge Pq) \rightarrow (P(p \wedge q) \vee P(p \wedge Pq) \vee P(Pp \wedge q))$

Ax9. $Gp \rightarrow Fp$ (futuro infinito)

Ax10. $Hp \rightarrow Pp$ (pasado infinito)

Ax11. $Fp \rightarrow FFP$ (densidad)

Los Ax0-Ax4 constituyen el sistema mínimo Kt^5 . Con el Ax5 y el Ax6 tiene modelos transitivos. Esto más Ax7 o Ax7' y Ax8 o Ax8' es Cl^6 , un sistema de tiempo lineal. A cualquiera de estos sistemas se les puede añadir la infinitud o la densidad.

En cuanto a la semántica, definimos un modelo de la lógica temporal como una estructura $\langle T, <, v, >$, donde

1. $T \neq \emptyset$ es un conjunto de momentos.

2. $< \subseteq T^2$, es la relación de ulterioridad, o relación antes/después, y tiene las propiedades que corresponden a cada sistema, siendo en todo caso irreflexiva.

3. v es una función de evaluación que asigna el valor verdadero o falso a cada fórmula bien formada (fbf) en un momento (siendo F el conjunto de fbfs, $v: FxT \rightarrow \{1, 0\}$). A partir de aquí se pueden obtener los subconjuntos de momentos $V(p) \subseteq T$ en los que una fórmula atómica es verdadera. La función de evaluación cumple las siguientes condiciones, para cualesquiera $t, s \in T$, $A, B \in F$ y variable proposicional p :

(i) $v(p, t) = 1$ o $v(p, t) = 0$

(ii) $v(\neg A, t) = 1$ syss $v(A, t) = 0$

(iii) $v(A \rightarrow B, t) = 1$ syss $v(A, t) = 0$ o $v(B, t) = 1$

⁵ Prior (1967).

⁶ De Cocchiarella.

$$(iv) \ v(FA, t) = 1 \text{ syss } \exists s \in T: t < s, \ v(A, s) = 1$$

$$(v) \ v(PA, t) = 1 \text{ syss } \exists s \in T: s < t, \ v(A, s) = 1$$

Decimos que una fórmula A es LT-válida (y lo representamos como $\models_{LT} A$) si, para todo $t \in T$, $v(A, t) = 1$.

3. Lógica epistémica

El lenguaje para la lógica epistémica L_{S5} para h agentes, se obtiene de añadir a un lenguaje para la lógica proposicional clásica, los operadores monarios K ("saber que") y M ("considerar posible que"). Así tenemos que si A es una fórmula bien formada, también lo son $K_i A$ ("el agente i sabe que A ") y $M_i A$ ("el agente i considera posible que A "). Se cumple, además, la siguiente definición metalingüística: $M_i A =_{df} \neg K_i \neg A$.

En el sistema axiomático, las reglas de deducción son Modus Ponens, Sustitución y Generalización de K .

Los axiomas de una lógica epistémica $S5$ pueden ser los siguientes:

Ax0. Todas las tautologías de la lógica clásica de proposiciones

Ax1. $K_i p \rightarrow p$ (Para $i = 1, \dots, h$)

Ax2. $\neg K_i p \rightarrow K_i \neg K_i p$ (Para $i = 1, \dots, h$)

En cuanto a la semántica, definimos un modelo de la lógica epistémica $S5$ como una estructura $\langle M, R_1, \dots, R_h, v \rangle$, donde

1. $M \neq \emptyset$ es un conjunto de estados o mundos epistémicamente posibles

2. $R_i \subseteq M^2$ ($i = 1, \dots, h$), son relaciones de accesibilidad entre mundos o estados epistémicos, una para cada agente, siendo, además, relaciones de equivalencia.

3. v es una función de evaluación que asigna el valor verdadero o falso a cada fórmula bien formada (fbf) en un mundo o estado epistémico (siendo F el conjunto de fbfs, $v: FxM \rightarrow \{1, 0\}$). A partir de aquí se pueden obtener los subconjuntos de mundos $V(p) \subseteq M$ en los que una fórmula atómica es verdadera. La función de evaluación cumple las siguientes condiciones, para cualesquiera $m, n \in M$, $A, B \in F$ y variable proposicional p :

$$(i) \ v(p, m) = 1 \text{ o } v(p, m) = 0$$

$$(ii) \ v(\neg A, m) = 1 \text{ syss } v(A, m) = 0$$

$$(iii) \ v(A \rightarrow B, m) = 1 \text{ syss } v(A, m) = 0 \text{ o } v(B, m) = 1$$

$$(iv) \ v(K_i A, m) = 1 \text{ syss } \forall n \in M: m R_i n, \ v(A, n) = 1$$

Decimos que una fórmula A es $S5$ -válida ($\models_{S5} A$) si, para todo $m \in M$, $v(A, m) = 1$.

4. Sistemas de lógica temporal y epistémica

4.1. Lógica temporal epistémica mínima

Engelfriet⁷ presenta un sistema mínimo para la lógica temporal epistémica (TEM), en donde poder expresar los cambios que experimenta el conocimiento de un agente a lo largo del tiempo, aunque este sistema no nos permite hablar del conocimiento que un agente tiene acerca del futuro y el pasado (por tanto, en el sistema de Engelfriet nos encontraremos operadores epistémicos bajo el alcance de operadores temporales, pero nunca operadores temporales que aparezcan bajo el alcance de operadores epistémicos).

Para ello utiliza un sistema de tiempo lineal transitivo⁸ y un sistema epistémico S5. De este modo, el sistema axiomático para TEM será el del tiempo lineal transitivo, añadiendo que cualquier fórmula demostrable en un sistema epistémico S5 también lo es en este sistema (es decir, que para toda fórmula $A \in L_{S5}$, si $\vdash_{S5} A$ entonces $\vdash_{TEM} A$).

Por lo que respecta a la semántica, definimos la validez de una fórmula A respecto a la Lógica Temporal Epistémica Mínima (TEM), es decir $\models_{TEM} A$, empleando la misma estructura que para la lógica de tiempo lineal (con la propiedad transitiva en la relación $<$), añadiendo a las cláusulas de la función de evaluación que

$$\text{Si } A \in L_{S5}, \text{ entonces } \models_{TEM} A \text{ syss } \models_{LTA} A$$

4.2. La lógica temporal epistémica de Kraus y Lehmann

Kraus y Lehmann (1988) proponen un sistema lógico que nos permite considerar conjuntamente el conocimiento y la creencia. El lenguaje L_{KL} de la lógica de Kraus y Lehmann se obtiene de añadir a un lenguaje para la lógica proposicional clásica, los operadores monarios K (conocimiento), B (creencia), C (conocimiento común), D (creencia común), E (conocimiento conjunto) y I (creencia conjunta). Así, si A es una fórmula bien formada también lo son las siguientes: KA , BA , CA , DA , EA y IA . Encontramos también las siguientes definiciones metalingüísticas:

$$EA =_{df} K_1 A \wedge \dots K_h A \quad (\text{siendo } h \text{ el número total de agentes})$$

$$IA =_{df} B_1 A \wedge \dots B_h A \quad (\text{siendo } h \text{ el número total de agentes})$$

$$CA =_{df} A \wedge EA \wedge EEA \wedge EEEA \wedge \dots = \forall i \geq 0 E^i A$$

$$DA =_{df} IA \wedge IIA \wedge IIIA \wedge \dots = \forall i > 0 I^i A$$

Las reglas de derivación del sistema axiomático KL son Modus Ponens, Sustitución y Generalización de C: Si $\vdash A$, entonces $\vdash CA$.

Además, el sistema axiomático KL consta de los siguientes axiomas:

Ax0. Todas las tautologías de LCP

Ax1. $K_i(p \rightarrow q) \rightarrow (K_i p \rightarrow K_i q)$

⁷ Engelfriet (1996).

⁸ Es decir, que la relación de ulterioridad $<$ cumple la propiedad transitiva: $<$ es transitiva si $\forall t, u, v \in T (t < u \wedge u < v \Rightarrow t < v)$.

- Ax2. $K_i p \rightarrow p$
 Ax3. $\neg K_i p \rightarrow K_i \neg K_i p$
 Ax4. $C(p \rightarrow q) \rightarrow (Cp \rightarrow Cq)$
 Ax5. $Cp \rightarrow K_i p$
 Ax6. $Cp \rightarrow K_i Cp$
 Ax7. $C(p \rightarrow Ep) \rightarrow (p \rightarrow Cp)$
 Ax8. $B_i(p \rightarrow q) \rightarrow (B_i p \rightarrow B_i q)$
 Ax9. $\neg B_i(p \wedge \neg p)$
 Ax10. $D(p \rightarrow q) \rightarrow (Dp \rightarrow Dq)$
 Ax11. $Dp \rightarrow Ip$
 Ax12. $Dp \rightarrow IDp$
 Ax13. $D(p \rightarrow Ip) \rightarrow (Ip \rightarrow Dp)$
 Ax14. $K_i p \rightarrow B_i p$
 Ax15. $B_i p \rightarrow K_i B_i p$
 Ax16. $Cp \rightarrow Dp$

Respecto a la semántica, decimos que un modelo-KL es una estructura $\langle M, R_1 \dots R_h, S_1 \dots S_h, v \rangle$, donde:

1. M es un conjunto de estados ($M \neq \emptyset$).
2. $R_i \subseteq M \times M$ son relaciones de accesibilidad respecto al conocimiento.
3. $S_i \subseteq M \times M$ son relaciones de accesibilidad respecto a la creencia.

Tales que:

- (a) R_i son relaciones de equivalencia.
- (b) S_i son seriales⁹.
- (c) $S_i \subseteq R_i$
- (d) $\forall m, n, o \in M (mR_i n \wedge nS_i o \Rightarrow mS_i o)$

4. Siendo F el conjunto de fbfs, $v: F \times M \rightarrow \{1, 0\}$, que cumple las siguientes condiciones, para cualesquiera $m, n \in M$, $A, B \in F$ y variable proposicional p :

- (i) $v(p, m) = 1$ o $v(p, m) = 0$
- (ii) $v(\neg A, m) = 1$ syss $v(A, m) = 0$
- (iii) $v(A \rightarrow B, m) = 1$ syss $v(A, m) = 0$ o $v(B, m) = 1$
- (iv) $v(K_i A, m) = 1$ syss $\forall n \in M: mR_i n, v(A, n) = 1$
- (v) $v(B_i A, m) = 1$ syss $\forall n \in M: mS_i n, v(A, n) = 1$
- (vi) $v(EA, m) = 1$ syss $\forall n \in M: mR_i n, v(A, n) = 1$ y ... y $\forall n \in M: mR_h n, v(A, n) = 1$
- (vii) $v(IA, m) = 1$ syss $\forall n \in M: mS_i n, v(A, n) = 1$ y ... y $\forall n \in M: mS_h n, v(A, n) = 1$

⁹ Decimos que una relación S es serial si $\forall m \in M, \exists n \in M: mS_n$. En este caso también se cumple que las S_i son transitivas (ver nota 8) y euclídeas (una relación S es euclídea si $\forall m, n, o \in M (mS_n \wedge mS_o \Rightarrow nS_o)$).

(viii) $v(CA, m) = 1 \text{ syss } \forall n \in M: mR^*n, v(A, n) = 1 \text{ (también } mR^*m)$

(ix) $v(DA, m) = 1 \text{ syss } \forall n \in M: mS^*n, v(A, n) = 1$

Decimos que una fórmula A es KL-válida ($\models_{KL} A$) si, para todo $m \in M$, $v(A, m) = 1$.

Kraus y Lehmann extienden esta lógica para incluir el tiempo mediante la adición de nuevos operadores temporales al anterior lenguaje. Aquí nos interesa, concretamente, el operador temporal $\circ$ (momento siguiente). Así, si A es una fórmula bien formada, también lo será $\circ A$.

Cuando hacemos referencia a cómo cambian las creencias a lo largo del tiempo, resulta interesante distinguir, a juicio de Kraus y Lehmann¹⁰, entre dos tipos básicos de creencia:

a) Creencia en sentido débil (creencia como hipótesis). Es un tipo de creencia que podemos interpretar como “estar dispuesto a apostar”. Así, por ejemplo, si el agente i cree que mañana lloverá, esto significa, operacionalmente, que mañana llevará un chubasquero consigo cuando salga de su casa. Si su creencia no se confirma y mañana no llueve, el problema no es muy grave, sólo que i se sentirá quizás incomodo al tener que estar cargando el chubasquero de un lado para otro. Respecto a este tipo de creencia, parece razonable afirmar que si uno cree (hoy) que mañana será el caso que A, uno también cree (hoy) que mañana seguirá creyendo que A es el caso. En relación a esta propiedad podemos introducir el siguiente axioma:

Ax17. $B_i \circ A \rightarrow \circ B_i \circ A$

Pero en realidad este axioma no dice nada sustancial sobre las creencias futuras, sino que hace referencia fundamentalmente al estado actual de las creencias. Más interesante sería poder expresar que si un agente i cree algo, entonces mantendrá su creencia hasta que sepa que es falsa. El medio para decir esto es:

Ax18. $B_i \circ A \rightarrow \circ B_i A \vee \circ K_i \neg A$

El siguiente axioma es una versión débil del Ax18:

Ax19. $B_i \circ A \rightarrow \circ B_i A \vee \circ B_i \neg A$

La interpretación de este axioma es que uno deja de creer en algo cuando uno cree que es falso.

b) Creencia “seria”. Es un tipo de creencia que uno no está dispuesto a abandonar, dada su posición central en nuestro sistema de creencias. Se trata de creencias acerca de cosas que sabemos que son verdaderas o que no pueden ser probadas como falsas. Sería, en fin, un tipo de creencia próxima a la creencia religiosa. Así, respecto de este segundo sentido de creencia, si un agente i cree que algo será el caso en el futuro, él sabe que en el futuro no descubrirá que tal cosa no es el caso. Por tanto, el siguiente axioma parece razonable en relación a la creencia “seria”:

¹⁰ Kraus y Lehmann (1988), pp. 166-7.

$$\text{Ax20. } B_i \circ A \rightarrow K_i \circ \neg K_i \neg A$$

Así, si el agente i cree que en el futuro será el caso que A , puesto que no puede encontrar en el futuro que estaba equivocado, tiene toda la razón en perseverar en sus creencias, y así podemos aceptar que:

$$\text{Ax21. } B_i \circ A \rightarrow \circ B_i A$$

El Ax20 es derivable a partir del Ax21, y así podemos abandonar el Ax20. Pero Kraus y Lehmann, siguiendo en esto a Fagin y Halpern, consideran que el Ax21 nos aleja del objetivo de capturar nociones de creencia realistas. Seguramente, nos dicen, el verbo “creer”, en su uso cotidiano, se emplea con un significado próximo al de “estar dispuesto a apostar” (creencia como hipótesis).

En cualquier caso, como subrayan los mismos autores, continúa abierto el problema de encontrar una semántica adecuada para estos sistemas.

4.3. Combinación de lógica temporal y epistémica: desarrollos sintácticos y semánticos.

En el apartado 4 hemos visto un sistema axiomático de lógica temporal epistémica mínima y su correspondiente semántica. La capacidad de expresión de dicha lógica es muy limitada, pues no nos permite expresar los conocimientos que un agente tiene acerca del futuro y el pasado. Resulta interesante, entonces, estudiar diversos desarrollos que nos permitan mejorar el poder expresivo de la lógica temporal epistémica en el sentido apuntado.

Para conseguir esto, en el plano semántico podemos partir de un conjunto $M \neq \emptyset$ de mundos posibles o alternativas epistémicas y un conjunto de instantes $T \neq \emptyset$. Además, parece interesante comenzar distinguiendo, con respecto a cada instante $t \in T$, dos tipos de alternativas epistémicas diferentes:

a) Las alternativas epistémicas que un agente i considera posibles en un instante t respecto a dicho instante t . Llamémoslas $M_{i(t,t)}$. Así tenemos que, en cada instante t , $M_{i(t,t)} \subseteq M$, dándose una relación de equivalencia R_i entre los elementos de $M_{i(t,t)}$.

b) Las alternativas epistémicas que un agente i considera posibles en un instante t respecto a un instante anterior/posterior s . Este conjunto de alternativas epistémicas para cada instante t respecto de otro instante anterior o posterior s sería entonces $M_{i(t,s)} \subseteq M$. También entre los elementos de $M_{i(t,s)}$ habría de establecerse una relación de equivalencia Q_i .

Así las cosas, las alternativas epistémicas de $M_{i(t,t)}$ nos permitirían evaluar fórmulas del tipo $K_i A$, $M_i A$ para cada agente i en cada instante $t \in T$. A su vez, respecto a las alternativas epistémicas de $M_{i(t,s)}$ evaluaríamos las fórmulas del tipo $K_i P A$, $K_i H A$ (si $s < t$) y $K_i F A$, $K_i G A$ (si $t < s$).

¿Qué ocurre con la evaluación de las fórmulas que son meramente temporales? Las fórmulas del tipo $P A$, $H A$, $F A$, $G A$, donde A no tiene ninguna ocurrencia de operadores epistémicos, no pueden ser evaluadas en

relación a las alternativas epistémicas (que son siempre relativas a cada agente), sino en función del “mundo (estado) actual” en cada instante¹¹. Llamemos ρ_s al “estado actual” en el instante s . Entonces habrá de cumplirse que $\rho_s \in M_{i(t,s)} \cap M_{i(s,s)}$. Que $\rho_s \in M_{i(t,s)}$ nos asegura que se cumple que $K_i FA \rightarrow FA$ y $K_i GA \rightarrow GA$ (si $t < s$), y que $K_i PA \rightarrow PA$ y $K_i HA \rightarrow HA$ (si $s < t$). A su vez, que $\rho_s \in M_{i(s,s)}$, nos asegura que se cumple $K_i A \rightarrow A$ en todo instante $s \in T$. Lo que de hecho ocurra en el “estado actual” en un determinado instante puede que no sea conocido por un agente i , pero al menos ha de ser considerado posible por dicho agente. Serían válidas, pues, fórmulas como: $A \rightarrow M_i A$, $FA \rightarrow M_i FA$, $GA \rightarrow M_i GA$, $PA \rightarrow M_i PA$, $HA \rightarrow M_i HA$.

Por su parte, las fórmulas en las que encontramos operadores epistémicos bajo el alcance de operadores temporales (como $PK_i A$, $HK_i A$, $FK_i A$, $GK_i A$) se evaluarían, en cada instante $t \in T$, en función del valor de la fórmula $K_i A$ en alguno (o todos) los instantes anteriores/posteriores a dicho instante t .

El hecho de que ρ_s (así llamamos al “estado actual” en el instante s) sea un elemento de $M_{i(t,s)}$, es decir $\rho_s \in M_{i(t,s)}$, hace que en todo instante $t \in T$ sea válida la fórmula $PK_i A \rightarrow PA$ (si $s < t$) y $FK_i A \rightarrow FA$ (si $t < s$).

Finalmente, si introducimos la cláusula: $M_{i(s,s)} \subseteq M_{i(t,s)}$, para todo $t, s \in T$ tal que $t < s$, hacemos válida la expresión siguiente: $K_i FA \rightarrow FK_i A$. Con esto hacemos que en todo momento $s \in T$ el agente i conserve los conocimientos que en el pasado había establecido en relación a ese instante s , pudiendo también aumentar tal conocimiento (pero nunca “perderse” una vez establecido). Esto es así porque con la cláusula anterior el número de alternativas epistémicas que un agente considera posibles en un instante t respecto a un instante posterior s , es igual o mayor que el número de alternativas epistémicas que el agente considera posibles en el instante s respecto a dicho instante s . Intuitivamente, la “pérdida” de alternativas epistémicas a lo largo del tiempo podría interpretarse como la eliminación o descarte de posibilidades a medida que aumenta la información (el conocimiento) del agente.

5. Conclusiones

Si la lógica epistémica resulta ser una herramienta útil para la reflexión acerca del conocimiento, su temporalización, para dar lugar a un sistema temporal epistémico bidimensional, aumenta extraordinariamente todas las ventajas que a tal respecto puede ofrecernos esta lógica. En efecto, no parece que el conocimiento de los agentes (ya sea un agente humano, un cognoscente ideal, un grupo de individuos, un computador, etc.) sea algo que permanezca

¹¹ Tal “estado actual” se determinaría, en cada instante, en relación a un observador que se sitúa fuera del mundo.

inalterable a lo largo del tiempo, sino que es susceptible de incrementarse a medida que, por ejemplo, se dispone de nueva información. Por otra parte, el tratamiento formal nos permite percatarnos de las sutiles diferencias que se dan entre el conocimiento de un agente en relación al momento en que se encuentra, y su conocimiento en relación a un momento anterior o posterior, diferencias éstas que resultan del todo relevantes en el transcurso de muchos análisis.

A lo largo de este trabajo hemos podido dar cuenta de algunas de las principales aportaciones que se han realizado en la combinación de lógica temporal y epistémica, ofreciendo además algunos desarrollos sintácticos y semánticos que pueden contribuir a configurar sistemas de lógica temporal y epistémica que dispongan de una semántica apropiada y que reflejen el conocimiento y sus propiedades en relación al tiempo de una manera satisfactoria.

BIBLIOGRAFÍA

- Engelfriet, Joeri: "Minimal Temporal Epistemic Logic", *Notre Dame Journal of Formal Logic*, Vol.37, nº 2, Spring 1996.
- Finger, Marcelo y Gabbay, Dov: "Combining Temporal Logic Systems", *Notre Dame Journal of Formal Logic*, Vol.37, nº 2, Spring 1996.
- Gabbay, Dov, Hodkinson, Ian y Reynolds, Mark: *Temporal Logic*, Oxford University Press, 1994.
- Kraus, Sarit y Lehmann, Daniel: "Knowledge, Belief and Time", *Theoretical Computer Science*, 58, 1988.
- Meyer, J.-J. Ch. y van der Hoek, W: *Epistemic Logic for AI and Computer Science*, Cambridge University Press, 1995.
- Prior, Arthur: *Time and Modality*, Oxford University Press, 1967.

José Rafael Herrera González
Universidad de La Laguna
e-mail: rahego@ull.es

Margarita Vázquez Campos
Universidad de La Laguna
e-mail: mvazquez@ull.es

THE USUAL ARGUMENT FOR PROPOSITIONS

ANDREA IACONA

Università del Piemonte Orientale (Vercelli)

ABSTRACT: When philosophers talk about propositions they refer to entities that enjoy certain features: language-independence, mind-independence, structure, truth conditions different from those of sentences. Many philosophers think that the notion of such an entity is intuitively plausible. Some of them appeal to our ordinary understanding of sentences containing that-clauses and to inferences hinging on these sentences. But the argumentative strategy based on this appeal is not a good strategy. If ordinary discourse shows that there are things to which that-clauses refer, these things don't enjoy any of the features enjoyed by the entities philosophers talk about.

KEYWORDS: Proposition, that-clause, language-(in)dependence, inference to the best explanation.

I

The word 'proposition' occurs rather often in philosophical discourse on language. Sometimes this word is used as if the reader should clearly know what it means, while other times its introduction is preceded by some preliminary consideration. In the latter case, the usual way of justifying the duty granted to the word is by appeal to everyday language and folk psychology. Particularly, a certain argumentative strategy seems to be in high favour with some philosophers. The line of argument starts from our ordinary understanding of sentences containing clauses of the form 'that so-and-so', and usually inferences like the following are brought in:

Tom believes that snow is white;
therefore, there is something that Tom believes.

Tom believes that snow is white;
that snow is white is true;
therefore, Tom believes something true.

Tom believes that snow is white;
John doubts that snow is white;

therefore, John doubts what Tom believes.

Apparently, in the first case the conclusion is obtained by existential generalization over what seems to be the “part” of the meaning of the first sentence expressed by the clause ‘that snow is white’. In other words, the inference presupposes that the sentence ‘Tom believes that snow is white’ says something of the form: ‘Tom believes x ’. The same reading of this sentence is needed in the second case, where the application of the predicate ‘is true’ to the clause ‘that snow is white’ and the following existential generalization seem to suggest that the x believed by Tom is something of which we can say: ‘ x is true’. In the third case, again, the first sentence is taken to have the form ‘Tom believes x ’, the second sentence is taken to have the form ‘John doubts y ’, and the identity relation applies to these x and y . It seems plausible to regard the x occurring in the open sentence ‘Tom believes x ’ as ranging over objects, because if this is the case then the first conclusion is obtained by unproblematic objectual quantification, the second hinges on a genuine property attribution, and the third hinges on an attribution of identity between objects. Using a more technical jargon, that-clauses seem to occur in the sentences above as *singular terms* or *referential singular terms*. To say that an expression t occur in a sentence of the form $F(t)$ as a referential singular term is to say that t stands for an object x , and the sentence is true just in case the predicate F is true of x ¹. To say that the proper name ‘Tom’ occurs as a referential singular term in the sentence ‘Tom is bald’ is to say that there is a certain man, Tom, such that the sentence is true just in case the predicate ‘is bald’ is true of Tom. Similarly, the inferences above seem to presuppose that there is a *thing* that Tom believes (to wit, *that snow is white*), that is precisely what the clause ‘that snow is white’ refers to. As inferences like these are clearly valid, it is claimed, there must be things that are the referents of that-clauses, and these things must be abstract entities of some kind. Usually at this point the preliminary considerations let more refined philosophical reflections pass, and the writer outlines his own conception of the abstract entities in question. There are major disagreements among philosophers about what exactly these entities are. However, their views seem to overlap on some points, in that certain features are commonly attributed to these entities independently from this or that particular conception of them. If we list these basic features we get a general sketch of what kind of things, let us call them the *Ps*, the word ‘proposition’ refers to when is used in philosophy.

¹ See for instance Schiffer [1998].

(1) Ps are *language-independent* entities. The sentence ‘Snow is white’ belongs to English and the sentence ‘La neve è bianca’ belongs to Italian. But they both express a certain P. This P could equally have been expressed by sentences of a different language, or even by sentences of a possible language that doesn’t actually exist. Ps are expressed by sentences of different languages but they themselves belong to no language. Their existence doesn’t depend on their being expressed by sentences of this or that language, nor it even depends on their being expressed by sentences of *some* language, since there are Ps that have never been expressed.

(2) Ps are *mind-independent* entities. They are at the same time the contents conveyed by our sentences and the objects of our thoughts or mental states. We can have several “attitudes” towards Ps: assume them, believe them, doubt them, and so on. But their existence doesn’t depend on their actually being the object of our attitudes. They would have existed even if thinkers hadn’t existed.

(3) Ps are *structured* entities made out of “building blocks” of some kind, and their structure matches to some extent the structure of the sentences that express them. The P expressed by the English sentence ‘Snow is white’ must contain at least one constituent corresponding to the word ‘snow’ and one constituent corresponding to the word ‘white’. The way these constituents are related must correspond to the way the words ‘snow’ and ‘white’ are combined in the syntax of the sentence.

(4) Ps have *truth conditions*, and the way their truth conditions are specified is sensibly different from the way truth conditions of sentences are specified. The sentence ‘Snow is white’ has the truth condition it has in virtue of the fact that English speakers use the words occurring in it as they do. Had English speakers used the word ‘white’ in the way they actually use the word ‘green’, the truth condition of the sentence would have been different. Sentences have their truth conditions relatively to this or that language, i.e. they are true-in-English, true-in-Italian, and so on. Accordingly, their truth conditions are *contingent* on the practices of the speakers of the language they belong to. On the contrary, Ps have truth conditions independently of any language, therefore they have them *necessarily* or essentially.

When philosophers talk about ‘propositions’ they refer *at least* to Ps as defined by the features (1)-(4). At the same time, they seem to take for granted that the notion of P is intuitively clear or that it may easily be made clear by intuitive considerations. Accordingly, the appeal to our ordinary use of sentences containing that-clauses is usually intended to provide a *prima facie* intuitive evidence for Ps. The purpose of what follows is to show that the

argumentative strategy hinging on this appeal is not a good strategy. The question whether we are entitled to think that there *really* are things to which that-clauses refer is itself a controversial one. But the main point here is that even if we forget about this question and suppose we have an intuitive - at least *prima facie* - justification for treating that-clauses as referential singular terms, we don't have for that any intuitive justification for thinking that their supposed referents are Ps.

II

Let us bracket for a moment our philosophical theories and just consider what the non-philosopher would be willing to allow about the supposed referents of that-clauses. Whoever happened, in everyday conversation, to be asked about the meaning of the word 'proposition' as is used in philosophy, knows that what we are entitled to take out of intuitions is not very much. One thing any non-philosopher would surely be willing to allow is that the things we say, think, believe, assert, are in some way independent of the act of saying, thinking, believing, asserting them. To say that snow is white is for a certain speaker to do something, namely, to perform the act of saying a certain thing in a certain place and at a certain time. But *what* is said, *that snow is white*, is neither something that is done nor something that happens. We take what is said to be something that many other speakers have said (and many others probably will) in other places and at other times. Moreover, although the act of saying that snow is white is different from the act of assuming, believing, conjecturing or claiming that snow is white, we take the what is said, *that snow is white*, to be something that could also be assumed, believed, conjectured or claimed. If we take 'assumption' to stand for 'what is assumed', and similarly the other terms for what is believed, conjectured, claimed, rather than for the acts or mental states of believing, conjecturing, claiming, we can identify statements with assumptions, beliefs, conjectures, claims. Our way of thinking about the objects or contents of certain linguistic or mental acts as distinct from the acts themselves entitles us to group together different acts that we take to share the same object or content.

It is not clear whether there is something else to say besides this about the intuitive repertoire of the non-philosopher. However, the non-philosopher may be perfectly content with *this*, and the little she allows goes well with the inferences invoked by philosophers. As long as we follow the intuitive distinction between what is asserted, believed, assumed, and so on, and the act of asserting, believing, assuming it, there is a trivial sense in which we are entitled to say that *there are* things we assert, believe, assume, and so on. Suppose

Tom asserted that snow is white. We obviously are willing to allow that Tom asserted a certain thing, therefore that there is a certain thing Tom asserted. In similar fashion, when Tom is going to the cinema we are willing to allow that there is a certain thing Tom is doing, namely, going to the cinema. All is needed for inferences like that from “Tom believes that snow is white” to “There is something Tom believes” is to conceive what we say, state, assume, and so on, as distinct from the act of saying, stating, assuming it. Accordingly, if there is a claim that an appeal to inferences like this can support, it is the claim that we actually conceive this way the things we say, state, assume, and so on. However, this claim doesn’t seem to be particularly original nor in need of evidence, therefore an argument for it wouldn’t be very interesting. In fact, when philosophers appeal to inferences like those mentioned, what they want to get at in the end are the Ps. But the unoriginal claim and the claim that there are Ps are rather different claims.

Let us call *propositions* the things we’re able to isolate as objects or contents of our different linguistic or mental acts (or kinds of acts), i.e. as what we say, state, assume, believe, and so on. This informal definition may be regarded as intuitive in that it doesn’t require any sort of theorization but the simple identification of objects obtained by an intuitive distinction.² Propositions so defined don’t seem to enjoy any of the features (1)-(4) that characterize the Ps.

To begin with, they aren’t *language-independent* entities. *Nothing* in the evidence provided by ordinary language leads us to think they are. Suppose Tom tells us: “The show was terrific”. We take what Tom states, *that the show was terrific*, to be independent from his act of stating it. But this doesn’t seem to entail that we take what Tom states to be independent from the language in which it is stated i.e. to be something that would have existed even if English hadn’t existed or that could be expressed in any other language. It certainly doesn’t entail that we take what Tom states to be something that would have existed even if languages hadn’t existed at all. A normal competent speaker wouldn’t probably know what to say when asked about the language-independence of Tom’s statement. Our intuitions are rather silent on this point.

If there is *something* in the evidence provided by ordinary language, it rather goes in the opposite direction. A good reason *not* to think that proposi-

² The definition is not intuitive in the sense that accounts for some primitive idea of a proposition, in contrast with more philosophically elaborate notions. In fact there isn’t such an idea, since our intuitions don’t push us to “put together” the objects of our different linguistic and mental acts.

tions are language-independent entities is that we *identify* and *describe* them only in linguistic terms. The “material” for our identification and description of propositions is provided by the clauses of the form ‘that so-and-so’ we use to explicitly refer to them. These clauses are in turn nothing but the result of a nominalization of the sentences ‘so-and-so’ that they embed. Explicit reference to propositions essentially relies on the linguistic practices that entitle us to get clauses of the form ‘that so-and-so’ from the correspondent sentences ‘so-and-so’. One way in which clauses of the form ‘that so-and-so’ are usually introduced is by application of the truth predicate. Given any closed indicative sentence *s*, our language entitles us to infer its pleonastic equivalent ‘It is true that *s*’. From

(1) Snow is white

we are entitled to infer

(2) It is true that snow is white.

Both (1) and (2) seem to have the form ‘*a* is *F*’. (1) says that the predicate ‘is white’ is true of the referent of the word ‘snow’, that is, snow. (2) seemingly says that the predicate ‘is true’ is true of the referent of the complex expression ‘that snow is white’, that is, the proposition *that snow is white*. The complex expression ‘that snow is white’, obtained by the nominalization of (1), is all we need to refer to the proposition that snow is white. This is to say that in inferring (2) from (1) we derive a sentence involving an explicit reference to a proposition from a sentence that doesn’t involve such a reference.

Another class of cases in which our linguistic practices license the introduction of clauses of the form ‘that so-and-so’ is that of indirect speech reports. Suppose (1) is uttered by Tom. Then

(3) Tom said: ‘snow is white’

is a direct report of Tom’s utterance. The indirect counterpart of (3) is

(4) Tom said that snow is white.

(3) involves explicit reference to the sentence ‘snow is white’, while

(4) involves explicit reference to the proposition that snow is white. A sentence of the form ‘*a* said that *p*’ counts as indirect counterpart of a sentence of the form ‘*a* said: ‘*s*’ just in case *p* is either the same sentence as *s* or it is a sentence obtained by paraphrasing *s*. In the latter case, one or more expressions occurring in *s* may be replaced with other words taken to be equivalent to them. The replacement is both a matter of interlinguistic or homophonic *translation* and a matter of *context adjustment*. If we use the term ‘interpretation’ to include both translation and context adjustment, we may say that indirect speech reports differ from direct speech reports in that they embed in

that-clauses the interpretations of the utterances directly reported. Both in the case of the truth-predicate and in the case of indirect speech reports, the step from sentences that don't contain explicit reference to propositions to sentences that do contain such reference may be accounted in terms of introduction of that-clauses embedding either sentences occurring in the first class of sentences or sentences that count as interpretations of them. And since we identify and describe propositions by means of the that-clauses we use to explicitly refer to them, we get propositions out of the sentences whose interpretations we embed in these that-clauses. This is to say that we get the proposition *that s* out of the sentence *s'* just in case *s* is the same sentence as *s'* or it counts as an interpretation of *s'*.

The step from sentences to Ps is usually taken to be a step from the level of the linguistic form given to a certain content to the deeper level of the content itself. According to a traditional picture, when we refer to Ps we use our language to pick up objects that are in some sense *outside* language. We use the expression 'that snow is white' to pick up a certain P more or less in the same way in which we use the expression 'Bill Clinton' to pick up Bill Clinton. A corollary of this picture is the idea that translation is warranted by the expression of a common content. That is, the English sentence 'Snow is white' and the Italian sentence 'La neve è bianca' are translatable *because* there is a certain language-independent object to which they both are related in some way. However, this picture is not justified in the case of propositions. Ordinary discourse alone doesn't push us as far as to think that propositions are objects outside language in the same sense in which Bill Clinton is, and we have reason not to think that they are. Our ways of identifying and describing Bill Clinton are to a large extent language-independent, that is, they can't be extracted by a simple analysis of our linguistic practices involving the expressions 'Bill Clinton' or 'The president of United States'. On the other hand, all we can say about the proposition *that s* is implicitly contained in our linguistic practices involving the expression that we use to refer to it, namely, the singular term 'that *s*'. And since the singular term 'that *s*' is obtained by nominalization of a sentence *s*, the identification of the proposition *that s* seems to essentially depend on the identification of *s*. This is not quite the same thing as to claim that propositions are language-dependent entities *because* our criteria of identification for them are exclusively linguistic. Perhaps an argument from the exclusively linguistic nature of our criteria of identification for propositions to their language-dependence wouldn't be very compelling. At the same time, however, the simple fact that we only have linguistic criteria of identification for propositions is clearly a good reason against their language-

independence. In order to rightfully claim that propositions are language-independent entities one would need some language-independent criteria of identification, that is exactly what we lack.

III

As we are not entitled to think that propositions would have existed if languages hadn't existed, all the more reason why we're not entitled to think that propositions would have existed if *thought* hadn't existed. Just as our ordinary discourse doesn't presuppose the language-independence of the things we say, state, assume, conjecture, it doesn't presuppose their *mind-independence*. Suppose Tom says: "Do you know what I thought? That it would be nice to go to the cinema tonight." Tom thought *that it would be nice to go to the cinema*, something that is independent of its being thought by Tom. But this doesn't seem to entail that this thing, *that it would be nice to go to the cinema*, would have existed even if thinkers like Tom hadn't existed. If all we are entitled to say is that propositions can be thought, assumed, conjectured by more than one person, they are no more mind-independent than sentences are. The proposition that it would be nice to go to the cinema isn't mind-independent in a deeper sense than the sentence 'it would be nice to go to the cinema' is. Just as we have no reason to take for granted the philosophical picture according to which propositional contents are objects "out there" that we pick up with our accidental sentential "dress", we have no reason to take seriously the metaphor of "grasping" these contents, that goes well along with that picture.

The previous considerations about language-independence also trivialize the question whether propositions have *structure*. Since our only way of identifying propositions is by looking at sentences embedded in that-clauses, there is not much to say about propositional structure. If to claim that propositions are structured entities is to claim that they are entities made out of language-independent building blocks of some kind whose combination happens to match the combination of the words occurring in the sentences that express them, we are not entitled to say that propositions are structured entities. We can't individuate the building blocks of the proposition *that snow is white* in a way that is to some extent independent from a simple analysis of the sentence 'Snow is white'. On the other hand, if claiming that propositions are structured entities amounts to claiming that they have *exactly the same* structure as sentences, we may confidently say that propositions are structured entities. But we don't gain much in holding the uninteresting claim that the proposition

that snow is white and the sentence 'snow is white' have the same structure in this sense.

The fourth feature of the Ps is the way their *truth conditions* are specified, which is taken to be substantively different from the way truth conditions of sentences are specified. The difference would amount to the fact that (a) truth-conditions of sentences are specified in terms of contingent biconditionals, because they depend on actual linguistic practices, while (b) truth-conditions of Ps are specified in terms of "essential" or necessary biconditionals, because they don't depend on any linguistic practice. There are reasons to doubt both that (a) is true and that something like (b) holds for propositions. One might be inclined to endorse (a) on the basis of considerations about linguistic counterfactuals. English speakers, one might say, could have used the words occurring in the sentence 'Snow is white' in a different way, and it seems that if they had,

(5) 'Snow is white' is true iff snow is white
could have come out false. However, whether linguistic counterfactuals make (5) false depends on the way we describe linguistic counterfactuals themselves, and the way we describe them depends in turn on what we take a language like English to be. If we think that what counts as relevant to the identification of a language are syntactic considerations only, we take the sentence 'Snow is white' to be a purely syntactic entity capable of having different meanings in different possible worlds. If we instead think that both syntax and semantics as relevant to the identification of a language, we take the sentence 'Snow is white' to be defined both by its syntax and its actual meaning. In the first case we account for the counterfactual cases in terms of possible worlds in which the very same sentence has a different meaning. Then (5) comes out false in some of these world and therefore contingently true. In the second case the sentence 'Snow is white' keeps its actual meaning across worlds. Then (5) comes out true in all possible worlds and therefore necessarily true. The question whether biconditionals like (5) are necessary or not is rather idle, and should perhaps be replaced by the question of which of the two ways of treating languages is preferable. There are good reasons to favor the second way. First, it is intuitively plausible to think about English as a language defined both by its syntax and its semantics, and to regard a language with different semantics as a language different from English. Second, our notion of the truth-condition of a sentence seems to be inevitably bound to our understanding of the actual meaning of that sentence. When we think about the truth conditions of the sentence 'Snow is white' we think about possible states of affairs in which snow would have been white, i.e. possible worlds in which the

sentence *as we understand it* is true, no matter what the speaker do or whether they exist in these worlds. On the other hand, if we adopt the first solution we are committed to the claim that the truth condition of a sentence *s* in a certain possible world *w* is always relative to the meaning of *s* in *w*, where the meaning of *s* in *w* depends on the linguistic practices of the speakers of *w*. But there are complications to work out with this claim. There are possible worlds in which 'Snow is white' doesn't belong to the linguistic repertoire of any speaker, and possible worlds in which it is used by speakers in more than one identifiable way³. Third, linguistic counterfactuals aren't really a problem for the second solution. The necessity of (5) might appear inconsistent with the truth of sentences like

'Snow is white' could have been true even if snow weren't white.

But there is no real inconsistency at issue. What the sentence says may be captured by "There could be a language *L* such that 'Snow is white' is true in *L* and snow is not white". That is, we can account for counterfactuals like that considered if we take them to involve an implicit quantification over languages⁴.

Independently of whether (a) is true, we have no reason to think that something like (b) holds for propositions, for we have no reason to think that propositions are language-independent entities. On the contrary, since propositions are essentially identified by means of sentences, their truth conditions are bound to be sensitive to the way those sentences are used. From this point of view, the question whether (5) is necessary or not is utterly irrelevant. If we adopt the second solution sketched above, then both truth conditions of sentences and truth conditions of propositions turn out to be specified by necessary biconditionals⁵. If instead we want to treat the meaning of our sen-

³ I owe this point to the reading of Casalegno [1997].

⁴ This solution seems to be consistent with the treatment of language variables given in Peacocke [1978].

⁵ Stephen Schiffer objected that this solution would be inconsistent with the claim that we don't need the Ps. If languages are regarded as abstract entities that have their semantic properties essentially, no matter what linguistic practices do or don't obtain, then *sentences* turn out to enjoy just the features that propositions have been denied: language-independence, mind-independence, structure, essential truth conditions. About language-independence, Schiffer may be right in pointing out that the second way of treating languages leads us to take the existence of a language to be independent from the linguistic practices of its speakers. Accordingly, if claiming that sentences are language-independent entities amounts to claiming that they exist independently from the linguistic practices of the speakers that use them, then sentences are language-independent entities (although it is a bit weird to say so). But the important sense in which Ps are supposed to be language-independent is that in which they are independent from *languages*. That is, although they can be expressed by sentences of different languages, they themselves belong to no language, and their existence doesn't depend on

tences as world-relative, propositions also turn out to be world-relative. The proposition *that s* is identified in a certain world *w* by means of the clause 'that *s*' which embeds the sentence *s* as is used by the speakers of *w*. Therefore, the truth condition of the proposition *that s* in *w* depends on the way *s* is used in *w*. It follows that the truth condition that a certain proposition has in certain possible world always coincides with the truth condition in that possible world of the sentence which expresses it. There is no substantive difference between truth conditions of propositions and truth conditions of sentences.

IV

As propositions don't share any of the features that characterize the Ps, the claim that there are propositions is rather different from the claim that there are Ps. The first uninteresting claim is implicitly contained in our ordinary use of that-clauses, while the second and more interesting claim goes far beyond the surface of that use. All ordinary discourse shows - if it does - is that there are *things* that we take to be what we say, state, assume, and so on. In order to get at Ps we need claim that these things have the features (1)-(4), that is just what ordinary discourse does *not* show. Surprisingly enough, philosophers seem to worry much more about the defense of the first trivial claim than about the justification of the step from this to the second claim.

The case for the Ps is sometimes presented in the form of an *inference to the best explanation*, as if the Ps were the things that better than any other explain our ordinary use of that-clauses. However, since ordinary use is relevant only as far as the claim that there are propositions is concerned, there is no explanatory gain in adding to it the claim that propositions are Ps. We may infer the sentence "There is something that Tom believes" from the sentence

their being given linguistic expression. Ps would have existed even if languages like English or Italian hadn't existed. Unlike Ps, sentences - however they are conceived - aren't obviously language-independent in this sense. Schiffer may also be right in saying that sentences conceived this way turn out to be mind-independent entities, in that they exist no matter what conceptual practices obtain. But my claim still holds that propositions need not be more mind-independent than sentences are. Similarly, sentences are obviously structured entities. But my point about structure is that there is no way of identifying the structure of propositions independently from the structure of the sentence used to refer to them. The same holds for truth conditions: however we want to conceive truth conditions of sentences, my claim is that there is no substantive difference between truth conditions of sentences and truth conditions of propositions. In substance, no matter what we want to say about sentences, my main point remains that propositions aren't more language-independent than sentences are, neither they enjoy some special features that sentences lack.

"Tom believe that snow is white" more or less in the same way in which we may infer the sentence "There is something that Tom is doing" from the sentence "Tom is going to the cinema". Similarly, from the conjunction of "Tom believes that snow is white" and "John believes that snow is white", we may infer the sentence "John believes what Tom believes" more or less in the same way in which from the conjunction of "Tom is going to the cinema" and "John is going to the cinema" we may infer "John is doing what Tom is doing". One need not postulate special properties of the thing Tom and John believe more than we need postulate analogue properties for the thing they do. There is no explanatory gain in supposing that *going to the cinema* is, say, an abstract entity which has the property of being independent from the fact that there are agents or from the fact that people go to the cinema. Similarly, there is no explanatory gain in supposing that propositions have the properties of the Ps. An inference to the best explanation is normally an abductive argument of the form: the supposition that p explains (better than the supposition that m , n , etc.) phenomena q , r , s ; therefore p . When p is about the existence of a certain object, the argument goes as follows: phenomena q , r , s may be explained if we suppose that there is a certain object x which has certain properties, say A and B ; therefore such object exists. The properties A and B are attributed to the supposed object x because they play some role in our explanation of q , r , s . Subsequently, it may be the case that we are able to identify x in a way that is to some extent independent from our explanation of q , r , s , thus we obtain a confirmation of our supposition. The independent identification of x may lead us to attribute other properties to it, say C and D , and in this way we acquire new knowledge about x . The argument for the Ps, however, doesn't conform to this pattern. The linguistic phenomena q , r , s said to be in need of explanation are the inferences hinging on that-clauses. The existence of an object x is said to be postulated in order to explain q , r , s . But this object x is characterized by properties - the features of the Ps - that play *no role at all* in the explanation of q , r , s . An attribution of properties that play no role in the explanation of q , r , s , to our object x would be justified only in case we had some other criterion of identification of x that is independent from the explanation of q , r , s . But we don't have such a criterion, since our only way of describing x is extracted from the same class of linguistic phenomena that x is supposed to explain. We have no way of describing x that is not reducible to "the object that plays such and such role in the explanation of q , r , s ", that is, we have no way of describing the objects that we postulate as referents of that-clauses that is not reducible to their being the referents of that-clauses.

REFERENCES

- Casalegno, Paolo , “Le proprietà modali della verità”, in *Logica e Teologia*, edited by A. Fabris, G. Fioravanti and E. Moriconi, Pisa, ETS, 1997, pp. 137-172.
Peacocke, Christopher , “Necessity and Truth Theories”, *Journal of Philosophical Logic*, 7, 1978, pp. 473-500.
Schiffer, Stephen , “Meaning and Concepts”, *Lingua & Stile*, 33, 1998.

Andrea Iacona

Università del Piemonte Orientale (Vercelli)

e-mail: iacona@cisi.unito.it





SORITES PARADOX FROM A LOGICAL POINT OF VIEW

H.J. MARRAUD

Universidad Autónoma de Madrid

ABSTRACT: Sorites paradox is considered so far as logical consequence is concerned. From this point of view the paradox threatens the transitivity of the consequence relation. Although transitivity (alongside with reflexivity and monotony) is a fundamental property of consequence, I argue that since its failure is due to the occurrence of vague predicates, a conservative revision of the concept of consequence, leading to a kind of substructural logic, will account for the paradox.

KEYWORDS: Sorites paradox, substructural logics, transitivity, vagueness.

Let us consider a typical case of the sorites puzzle:

One is few

If one is few, two are few

If two are few, three are few

.

.

.

If 9.999 are few then 10.000 are few

Therefore 10.000 are few

The first premise seems indisputably true -if some quantity at all is few, one is few; as for the others, conditional premises, they all take their plausibility from a general principle, expressed by the schema

(1) If n are few then $n+1$ are few.

This principle rests on the intuition that simply adding a single unit is not enough to pass from few to not few. The conclusion, however, seems unacceptable -although all this depends partly on the subject (grains of sand, millet seeds, people, etc.) and purpose of the conversation. Even if adding one unit does not make any significant difference, adding 9.999 units can do it. It is

worth noting that few lacks both compactness and the Löwenheim-Skolem property.

By “a logical approach” I mean one concerned with the concept of consequence rather than with the concept of truth; my interest here is with the effects of the sorites paradox on the current concept of logical consequence.

Depending on the chosen style of presentation (sequent calculus, natural deduction, Hilbert style calculus) there are several different formal versions of the above argument. These formulations do not have the same heuristic value as we shall see. Beginning with sequent calculus let me recast the above argument as

$$\begin{array}{l}
 \text{[axioms]} \quad \frac{}{\vdash \text{Few}(1)} \quad \frac{}{\text{Few}(1) \vdash \text{Few}(2)} \\
 \text{[Cut]} \quad \frac{}{\vdash \text{Few}(2)} \quad \frac{}{\text{Few}(2) \vdash \text{Few}(3)} \quad \text{[axiom]} \\
 \text{[Cut]} \quad \vdash \text{Few}(3) \\
 \vdots \\
 \vdots \\
 \text{[Cut]} \quad \frac{}{\vdash \text{Few}(9.999)} \quad \frac{}{\text{Few}(9.999) \vdash \text{Few}(10.000)} \text{[axiom]} \\
 \text{[Cut]} \quad \vdash \text{Few}(10.000)
 \end{array}$$

Here the crucial logical principle is the cut rule. Hence for this version of the paradox it seems that there are two possible analysis: either some premise (axiom) is false or cut is not an admissible rule in contexts like this. However, if we use natural deduction instead of sequent calculus, we get a formulation of the argument in which the logical principle at stake is *modus ponens*:

$$\begin{array}{ll}
 \text{Few}(1) & \text{[premise]} \\
 \text{Few}(1) \supset \text{Few}(2) & \text{[premise]} \\
 \text{Few}(2) & \text{[modus ponens]} \\
 \vdots & \\
 \vdots & \\
 \vdots & \\
 \text{Few}(9.999) & \text{[premise]} \\
 \text{Few}(9.999) \supset \text{Few}(10.000) & \text{[premise]} \\
 \text{Few}(10.000) & \text{[modus ponens]}
 \end{array}$$

One advantage of the sequent calculus is its distinction between structural rules, “pure” principles governing $\vdash$ in which no truth-functional connectives appear, and operational (also called logical) rules, principles governing the interaction of $\vdash$ with individual truth-functional connectives. Structural rules introduce general features of the consequence relation, features that are independent of the presence of any logical constants, while operational rules define the derivational use of logical constants. Thus a disagreement over

structural rules is a disagreement over the general concept of an argument -the question may be, e.g., whether the premises form just a set of formulae or a richer structure (a multiset, a sequence, etc.)- not over the meaning of logical constants. In natural deduction, on the other hand, the content of structural rules is expressed by the whole set of elimination/introduction rules. Therefore a sequential approach seems to have some additional heuristic value here.

Given the sequential analysis of the argument, there are two main accounts of the sorites paradox, as it has been said. According to the first, the argument is formally right but unsound; this entails the denial of the “major” premise, the schema (1). Thus it is held that in the argument occurs some false instance of (1), the problem being then to determine which. It is often claimed that typically we do not, and perhaps cannot, know where it is. In other words, as a result of the vagueness of the concept of few, we are unable to find out where is the false premise, although the conclusion being false and the argument valid, we know that there has to occur some false conditional premise. Thus the first, epistemic intended solution is logically conservative. This approach has some appealing consequences for the concept of sound argument. It is held that the above argument is valid but not sound; perhaps it is clearly so, even if we are unable to give a precise account of its not being sound. But if we agree both that ‘10.000 are few’ is false and that the preceding argument is (formally) valid, we have to conclude that at least one premise is false, though we cannot determine which. Let ‘If $k+1$ are few then $k+2$ are few’, $1 < k < 10.000$, the first (and possibly the sole) false premise in the argument. Then the shorter argument

One is few

If one is few, two are few

.

.

.

If k are few then $k+1$ are few

Therefore $k+1$ are few

will be both valid and sound, all its premises being true, although we would be unable to recognise this fact. While the main argument was valid and unsound and clearly so, the later argument is sound but not clearly so. It seems then that sound argument is a vague concept too. The point is that while the argument form shared by these arguments is thought to be valid, it will be of little use to draw conclusions, since we cannot be sure whether a particular instance of this form is sound or not. Hence vagueness, is transferred from the

truth/falsity of premises to the reliability/unreliability of whole arguments, rendering useless some valid argument forms. Indeed, the distinction unvalid vs. unreliably valid arguments is quite subtle.

Under the second approach the problem lies in the iterated use of the cut rule. The cut rule allows us to combine arguments into a longer argument, preserving validity, thus expressing the transitivity of reasoning. It is a well-known fact that vagueness and transitivity do not fit well; e.g.

When the identity relation is vague, it may seem intransitive; a chain of apparent identities may yield an apparent non-identity. Some sort of "counterpart" notion (...) may have some utility here.

S. Kripke, *Naming and Necessity*, fn. 18.

T. Hence it may seem reasonable to suggest that so far as few (like other concepts involved in similar arguments) is a vague notion, cut, an expression of the transitivity of the consequence relation, fails. The thesis is that *each* argument with the form

n are few

if *n* are few then *n*+1 are few

Therefore *n*+1 are few

is all right, but that they cannot be chained together, this being an effect of vagueness. Admittedly this looks a bit bizarre. Let us consider therefore a formal modelling in order to gain a better understanding of this proposal. Given a first-order language *L* and some supply of two-valued *L*-models **M**, let *f* be a function assigning to every sentence (set of sentences) *A* (*X*) of *L* a subset of **M** in some suitable systematic manner. Roughly speaking, *f*(*A*) gives us the meaning of *A*. Next define consequence, for the simplest case, as follows: for any two sentences *A* and *B* of *L*,

$A \models B$ iff for every model *M* $\varepsilon f(A)$, if $[A]^M = 1$ then $[B]^M = 1$.

(This has some affinity with McCarthy's circumscription device -cf. McCarthy [1980]). Let *A*, *B* and *C* be three *L*-sentences and suppose that for every *M* $\varepsilon f(A)$, either $[A]^M = 0$ or $[B]^M = 1$, and for every *M'* $\varepsilon f(B)$, either $[B]^{M'} = 0$ or $[C]^{M'} = 1$. Therefore both $A \models B$ and $B \models C$. Nonetheless it may be the case that for some *M''* $\varepsilon f(A) \cap \mathbf{M} - f(B)$, $[A]^{M''} = 1$ and $[C]^{M''} = 0$, so that not $A \models C$. Now to endorse the validity of such a principle like (2) $\text{Few}(n) \vdash \text{Few}(n+1)$ is to claim that in any $\text{Few}(n)$ -model in which this sentence holds, the sentence $\text{Few}(n+1)$ holds too. As $\text{Few}(n)$ -models and $\text{Few}(n+1)$ -models can be different, $\text{Few}(n+2)$ can fail in some such $\text{Few}(n)$ -model while being verified by every relevant $\text{Few}(n+1)$ -model.

Alternatively, a Kripke style semantics (similar to the current Stalnaker-Lewis semantics of conditional logics) can be used. Define a model for L as a quadruple $\langle W, R, f, V \rangle$ where W is a set (of possible worlds), $R \subseteq W^2$ is an alternativeness relation on W , f is a function from formulas into the power set of W , and V is an interpretation of the usual kind. Then define, for any formulae A and B of L ,

$A \models B$ iff for every admissible model $\langle W, R, f, V \rangle$ of L and $w \in W$, if
 $w \in f(A)$ and $V(A, w) = 1$ then $V(B, w) = 1$.

The selection function f works as a device to represent vagueness; the idea is roughly that when vague predicates occurs in a sentence A , not every model of L will count as a relevant model for A . Given a sentence A , you can think of $f(A)$ as the set of models in which the truth value of A , whether true or false, is clear. Hence it may be assumed that in absence of vague predicates two sentences pick the same models, i.e., $f(A) = f(B) = M$. So since transitivity failures are allegedly linked with vagueness, cut will work for precise statements.

Notice that though logically heterodox, this is not a many-valued approach preaching for a non-classical, many-valued logic within which we can refuse to accept the major premise without being committed to a sharp cut-off. On the contrary, here the major premise $Few(n) \vdash Few(n+1)$ is accepted. Moreover, when you embrace many-valuedness you are somehow committed to the thesis that truth comes in degrees, but M is a set of two-valued models.

Thus in general given a vague concept like few one has to reject cut. There are several versions of the cut rule in the literature; the following version is called *transitivity (lemma generation)* by Gabbay [1994]:

$$\frac{X, \vdash A \quad X, A \vdash B}{X \vdash B}$$

As a matter of human practical reasoning (as opposed to purely deductive reasoning), we could still consider some kind of approximate reasoning lead by the principle “in vague contexts, the reliability of an argument decreases with every application of cut”. As Makinson [1994, p.43] writes:

Cut may also be seen as expressing a determination not to allow the length, intricacy or manner of derivation of a conclusion to reduce the freedom with which it is used as a lemma in further inference. There is no “diminution of usability” with distance from origins. Once inferred, a proposition may be called upon in conjunction with the original information, and anything obtained with their combined use will be implied by the original information. It is not surprising that most forms

of probabilistic inference fail cut, since probabilities of conclusions may diminish as chains of inference grow longer.

Perhaps schema (1) could be seen as a default, in the sense of Reiter [1980], something like

$$\frac{\text{Few}(n):\text{Few}(n+1)}{\text{Few}(n+1)}$$

This rule means, intuitively, that a ground instance of $\text{Few}(n+1)$ can be used as a premise if the corresponding instance of $\text{Few}(n)$ is proved and the corresponding instance of $\text{Few}(n+1)$ is consistent with everything believed. An obvious difficulty with this suggestion is that the argument we have been considering has a converse,

$$\begin{array}{l} \neg\text{Few}(10.000) \\ \text{If } \neg\text{Few}(10.000) \text{ then } \neg\text{Few}(9.999) \\ \cdot \\ \cdot \\ \cdot \\ \neg\text{Few}(2) \\ \text{If } \neg\text{Few}(2) \text{ then } \neg\text{Few}(1) \\ \text{Therefore, } \neg\text{Few}(1) \end{array}$$

So there will be a corresponding default,

$$\frac{\neg\text{Few}(n+1):\neg\text{Few}(n)}{\neg\text{Few}(n)}$$

conflicting with the first. In any case this is another story.

The stated thesis that vagueness and cut collide seems to lead to a deep revision of classical logics and semantics. To be more precise, since cut is a structural rule, it points towards a substructural logic, a logic sharing with classical logic the same set of operational or logical rules but whose structural rules are a proper subset of classical ones. Since structural rules introduce features of the consequence relation independent of the presence of any logical constant, it will be a logic with exactly the same operators that classical logic but with a different notion of an argument.

Despite appearances on the contrary, this is a minor departure from classical logic. For vagueness is the alleged reason to reject transitivity or cut, and the relevant kind of vagueness here affects predicates but not logical constants. It is to be remembered that cut-elimination usually fails in sequent systems with proper axioms; e.g., the system obtained adding to a suitable sequent formulation of classical first-order logic the axioms

$$\vdash t=t \qquad t=s, A[t] \vdash A[s].$$

where 's' and 't' stand by singular terms, 'A' is a wff, and 'A[t]' means that 't' occurs somewhere in 'A' and 'A[s]' means that 's' replaces 't' in 'A' in some occurrences.

For instance, the sequent

$$a=b, b=c, Pa \mid Pc$$

is easily proved using cut:

[axioms]	$a=b, Pa \vdash Pb$	$b=c, Pb \vdash Pc$
[left-monotony]	$a=b, Pa, b=c \vdash Pb$	$b=c, Pb, a=b, Pa \vdash Pc$
[left-permutation]	$a=b, b=c, Pa \vdash Pb$	$a=b, b=c, Pa, Pb \vdash Pc$
[cut]	$a=b, b=c, Pa \vdash Pc$	

but it cannot be proved without that rule.

If you work with a purely logical system S consisting in logical axioms like *Identity*

$$A \vdash A,$$

structural rules like *left weakening*

$$\frac{X \vdash Y}{X, A \vdash Y}$$

and operational rules like

$$\frac{X, A \vdash B, Y}{X \vdash A \supset B, Y} \quad \frac{X \vdash A, Z \quad Y, B \vdash W}{X, Y, A \supset B \vdash Z, W}$$

the cut rule will be eliminable, if it is a rule of S, or an admissible rule, if it is not; i.e., for every proof of a sequent S using cut there will be a cut-free proof of the same sequent. This fact expresses the deep and nice symmetries of logic allowing for normalization of proofs. When you go beyond, adding proper, extralogical axioms, to get a formal theory (e.g. elementary arithmetics) you *need* cut: either cut is no longer eliminable, if it was a rule of S, or it has to added, if it were not. If S is a sequent calculus not having the cut rule, $S' = S + \text{cut}$, and X is a set of proper axioms, then in general $\{X \vdash Y / X \vdash Y \text{ is } S\text{-provable}\} \subset \{X \vdash Y / X \vdash Y \text{ is } S'\text{-provable}\}$.

Now it has been claimed that if the proper axioms of your theory involve some vague concepts (like few) then cut has to be somehow restricted. Thus changes take place out of the purely logical core, in the treatment of formalized theories.

The interweave of transitivity (cut) and monotony deserves some commentary. The converse of cut is a restricted form of monotony, called *cautious monotony* by Gabbay [1985]. It is the condition

$$X \cup \{A_i / i \in I\} \vdash B \text{ whenever } X \vdash A_i \text{ for all } i \in I \text{ and } X \vdash B.$$

A consequence operation satisfying both cut and cautious monotony is called *cumulative*. These two conditions together ensure the stability of the reasoning

process, telling us that when we are reasoning we can safely accumulate our conclusions into our premises.

When we move from deductive reasoning to common sense reasoning, where uncertainty is ubiquitous, there is a loss of stability -as it is acknowledged by logicians working in nonmonotonic or defeasible logic. For instance, it can be verified that any cumulative consequence operation satisfying transitivity and well-behaved with respect to conjunction yields a finitary version of monotony. For this reason some systems of nonmonotonic reasoning, wishing to abandon finitary conjunctive monotony have to abandon transitivity too (cf. Makinson, 1994, pp.45s.).

Acknowledgements. The author would like to acknowledge the financial support of the Spanish DGICYT (project Modalidad y parcialidad en inteligencia artificial, PB96-1301-C05-C2).

REFERENCES

- Gabbay, D.M.: "Theoretical foundations for nonmonotonic reasoning in expert systems" en K.Apt (de.), *Logics and Models of Concurrent Systems*, pp. 439-457, Berlin, Springer Verlag, 1985.
- Gabbay, D.M.: "What is a Logical System" in D.M.Gabbay (de.), *What is a Logical System?*, pp. 179-216, Oxford, Clarendon Press, 1994.
- Gabbay, D.M., Hogger, C.J. and Robinson, J.A. (eds.): *Handbook of Logic in Artificial Intelligence and Uncertain Reasoning*. Oxford, Oxford University Press, 1994.
- Kripke, S.A.: *Naming and Necessity*, Cambridge, Mass., Harvard University Press, 1980.
- McCarthy, J.: "Circumscription-A form on nonmonotonic reasoning", *Artificial Intelligence* 28, 1980, 89-116.
- Makinson, D.: "General Patterns in Nonmonotonic Reasoning" in Gabbay, D.M. et al. (eds.), 1994, pp.35-110.
- Reiter, R.: "A logic for default reasoning", *Artificial Intelligence* 13, 1980, 81-132.

Hubert J. Marraud

Universidad Autónoma de Madrid

e-mail: hubertmarraud@uam.es

SOBRE LA RELACIÓN ENTRE LA PARADOJA DE RUSSELL Y LA DEL MENTIROSO

JOSÉ MARTÍNEZ FERNÁNDEZ¹
Universidad de Valencia

ABSTRACT: Priest has argued that most paradoxes of self-reference share a common underlying structure. Brendel, on the contrary, supports the view that mathematical and semantical paradoxes have differences that are philosophically relevant. This paper discusses both opinions and argues that the classification of paradoxes must be relativized to the kind of problem the solution is intended to solve.

KEYWORDS: Liar paradox, Russell paradox, self-reference.

1) Dos tipos de paradojas

A principios de este siglo proliferaron las paradojas ligadas a la autorreferencia, que afectaban a una amplia gama de conceptos. A pesar de la opinión de Russell, quien pensó que todas las paradojas se producen debido a que se introduce en su formulación un término definido circularmente, la opinión estándar considera que las paradojas se dividen en dos grupos fundamentales: las *paradojas matemáticas* (también llamadas paradojas de la teoría de conjuntos) y las *paradojas semánticas*. Esta opinión se remonta a Ramsey, quien en su artículo (Ramsey 1931), pp. 20s. distingue entre aquellas paradojas que envuelven nociones de teoría de conjuntos o lógicas, como la paradoja de Russell o la de Cantor, y aquellas otras en las que aparecían nociones semánticas, como las de “definible” o “verdad”. Las paradojas del segundo grupo (las semánticas) utilizan conceptos empíricos, como interpretación o denotación, y se producen porque se sostienen teorías incorrectas respecto a esos conceptos. Sólo muestran que algunas representaciones del lenguaje o el pensamiento ordinario no son correctas. Sin

¹ El autor quiere agradecer a la Fundación Caja Madrid el apoyo financiero prestado para la realización de este trabajo.

embargo, las paradojas matemáticas, al involucrar conceptos puramente matemáticos, como los de conjunto o cardinal, indican que algo no es correcto en la lógica utilizada en la ciencia, y su gravedad es mayor. Desde el desarrollo de la metamatemática, sabemos que las nociones semánticas tienen también una representación formal precisa, por lo que la idea de Ramsey de que conceptos como el de interpretación y verdad no son parte del lenguaje científico pierde validez. El objetivo de este trabajo es analizar algunas propuestas modernas respecto a la validez de la distinción entre las paradojas. Nos centraremos en las dos paradojas que podemos considerar paradigmáticas de cada grupo: la paradoja de Russell (el conjunto de todos los conjuntos que no se pertenecen a sí mismos) y la paradoja del mentiroso (la oración “esta oración es falsa”).

Una de las causas de que la distinción entre ambos grupos de paradojas haya sido adoptada como natural es que parece explicar el distinto tratamiento que han recibido la paradoja de Russell y la del mentiroso. La paradoja de Russell mostraba la inconsistencia de la teoría intuitiva de conjuntos, y con ella hacía peligrar la fundamentación de las matemáticas, así que la necesidad de encontrar una solución era imperiosa. La paradoja del mentiroso ponía en cuestión la teoría intuitiva de la verdad, pero esto no acarrea una dificultad para el lenguaje ordinario: el hablante ordinario resuelve las paradojas de un modo pragmático, considerándolas como oraciones singulares que pueden excluirse del ámbito de lo significativo. Cuando se trabaja en teoría de conjuntos se aceptan teorías axiomáticas que impiden la paradoja de Russell. Puede haber diferencias respecto a la teoría axiomática con la que resulta más conveniente trabajar, pero todas ellas resuelven la paradoja del mismo modo, y no hay discusión apenas respecto a si resuelven correctamente o no la paradoja. Sin embargo, aunque Ramsey mismo pensaba que las paradojas semánticas debían resolverse al modo de Russell, mediante restricciones del lenguaje, siguen proliferando y discutiéndose intentos de solución radicalmente distintos a la paradoja del mentiroso. Esto lleva a pensar que debe haber algún tipo de diferencia entre ambos conjuntos de paradojas.

2) La unidad estructural

Por paradoja se entiende un argumento que conduce a partir de premisas aparentemente verdaderas mediante pasos de deducción aparentemente válidos a una conclusión contradictoria. Una paradoja no es, propiamente hablando, una oración o una definición que dan lugar a una contradicción, sino un tipo de argumento. La distinción de Ramsey se centra en los conceptos que aparecen en las oraciones que provocan la paradoja, no en la estructura del argumento mismo. Aunque evidentemente los conceptos

involucrados son distintos, Priest (1984) sostiene que casi todas las paradojas revelan una misma estructura.

Concretamente, la mayor parte de las paradojas, tanto semánticas como matemáticas, pueden verse como instancias del siguiente esquema²: dadas dos propiedades ϕ y ψ y una función (posiblemente parcial) δ , entonces, si se cumple que

(1) el conjunto $w = \{x: \phi(x)\}$ existe y $\psi(w)$,

(2) si $x \subseteq w$ tal que $\psi(x)$, entonces:

(a) $\delta(x) \notin x$ (Condición de trascendencia),

(b) $\delta(x) \in w$ (Condición de clausura),

entonces tenemos una contradicción, dado que, por (1), $\psi(w)$ y, aplicando (2), $\delta(w) \notin w$ y $\delta(w) \in w$.

Veamos que las dos antinomias principales se pueden expresar según las condiciones de este esquema. La paradoja de Russell se produce al considerar, en un universo de conjuntos, como ϕ la propiedad $x \notin x$. Entonces el conjunto w es el conjunto de todos los conjuntos que no se pertenecen a sí mismos (al que llamaremos conjunto de Russell). Como ψ se considera una propiedad universal (como la de ser idéntico a sí mismo) y como δ a la función identidad. Sea $x \subseteq w$, entonces, si $x \in x$, $x \in w$ y, por definición de w , $x \notin x$. Por reducción al absurdo se sigue que $x \notin x$, luego, como δ es la identidad, se cumple la condición de trascendencia y, por definición de w , $x \in w$, luego se cumple la condición de clausura. Si suponemos que w existe, se cumple la condición (1) y, por tanto, se produce la contradicción.

La paradoja del mentiroso se formula considerando como universo los conjuntos de oraciones y como ϕ la propiedad de ser una oración verdadera. w es el conjunto de oraciones verdaderas, al que denotamos como V . Como ψ consideramos la propiedad de ser un conjunto definible de oraciones y como δ a la función que asigna a cada conjunto definible de oraciones una oración que afirma de sí misma que no pertenece a ese conjunto³. Si x es un conjunto definible de oraciones, representamos a la oración $\delta(x)$ como $\delta(x) = \text{"}\delta(x) \notin x\text{"}$.

² Pero no todas: la paradoja de Curry ("si esta oración es verdadera, entonces p ", donde p es una oración cualquiera) es, según Priest, un tipo de argumento diferente al desarrollado a partir del esquema si la interpretación del condicional no es la material.

³ El conjunto debe ser definible para asegurar que existe la oración. Si los portadores de la verdad son proposiciones, entonces a cada conjunto de verdades le corresponde la proposición que dice que ella no pertenece a ese conjunto, y la propiedad ψ ya no es relevante: puede elegirse como ψ una propiedad universal.

Sea $x \subseteq V$ un conjunto definible. Si $\delta(x) \in x$, entonces, como $x \subseteq V$, $\delta(x) \in V$. Por definición de $\delta(x)$, " $\delta(x) \notin x$ " $\in V$. La Convención T de Tarski, que recoge el sentido intuitivo del concepto de verdad, implica que es cierto el bicondicional " $\delta(x) \notin x$ " $\in V$ si, y sólo si, $\delta(x) \notin x$. Luego $\delta(x) \notin x$. Esto demuestra la condición de trascendencia. Además, como $\delta(x) \notin x$, por la convención T, " $\delta(x) \notin x$ " $\in V$, luego $\delta(x) \in V$, demostrando la clausura. La existencia de V implica, como en el caso anterior, la contradicción.

En ambos casos se demuestra que hay cierto conjunto que no puede existir: en uno el de los conjuntos que no se pertenecen a sí mismos, en otro el de las oraciones verdaderas. Estos argumentos pueden verse como variaciones del argumento de la diagonal de Cantor: supuesta la existencia del conjunto de todos los conjuntos que no se pertenecen a sí mismos o del conjunto de todas las verdades, se muestra la existencia de ciertos elementos (el mismo conjunto de Russell o la oración del mentiroso) que "diagonalizan" esa totalidad: deberían estar incluidos en la totalidad, pero por su misma definición no pueden pertenecer a ella⁴.

Si varias paradojas efectúan el mismo tipo de razonamiento, evitarlo sólo en algunos dominios no parece la mejor manera de evitar definitivamente las paradojas. De ahí que Priest proponga el *principio de solución uniforme*: a las paradojas que son del mismo tipo debe dárseles el mismo tipo de solución. Que sean del mismo tipo quiere decir que comparten una misma estructura, y que esa estructura produce la contradicción. En conclusión, dado que tanto la paradoja de Russell como la del mentiroso producen una contradicción del mismo modo, ambas deberían tener el mismo tipo de solución.

3) El mentiroso reforzado

Brendel (1992, pp. 49-54) retoma la distinción de Ramsey: aunque ya no tiene sentido distinguir si los conceptos son empíricos o no, sí hay una diferencia entre los conceptos. La paradoja de Russell sólo requiere para su formulación considerar objetos sintácticos del lenguaje de la teoría de conjuntos. Sin embargo, la paradoja del mentiroso requiere la consideración del concepto de verdad (o del de satisfacción) que relaciona el lenguaje con un cierto ámbito de interpretación: es decir, se utilizan conceptos que reflejan la relación del lenguaje con algo exterior a él, haciendo que sea parte del lenguaje la relación con el mundo.

⁴ Simmons (1993), pp. 16-44, presenta muchas formas de paradojas como instancias de un argumento generalizado de la diagonal, mostrando así de otro modo su igualdad estructural.

El problema es si esta distinción es relevante filosóficamente, de tal modo que favorezca un tratamiento diverso para ambas paradojas. Brendel cree que sí, puesto que sólo en la paradoja del mentiroso (y en las paradojas semánticas) se reproduce el problema de las paradojas reforzadas. Se llama paradoja reforzada del mentiroso a aquella que se crea al intentar expresar en el lenguaje objeto la solución teórica adoptada para la paradoja del mentiroso. Por ejemplo, si se declara la oración del mentiroso como carente de valor de verdad, al modo de Kripke, esto evita la contradicción y puede definirse un predicado verdad para un lenguaje objeto que contenga su propio predicado verdad. El problema es que puede considerarse la oración “esta oración es falsa o carece de valor de verdad”, que es contradictoria tanto si posee valor de verdad como si no lo posee. Si el lenguaje para el que definimos la verdad tiene los recursos suficientes para poder expresar la solución propuesta, es decir, para poder decir de una oración que carece de valor de verdad, entonces la oración reforzada puede escribirse y, de nuevo, se produce una paradoja. Todas las soluciones propuestas a la paradoja del mentiroso han sido víctimas de una u otra forma del mentiroso reforzado. Sin embargo, no sucede lo mismo con la paradoja de Russell, que no se ve sometida a este tipo de críticas.

En opinión de Brendel, esta diferencia respecto a las paradojas reforzadas es debida a una diferencia fundamental entre las teorías de los conceptos matemáticos y los semánticos: en el caso de los conceptos semánticos no podemos estar fuera del lenguaje que analizamos: la teoría es la metateoría. El lenguaje ordinario puede considerarse como matriz de todos los lenguajes, dado que siempre pueden traducirse a él: si queremos definir los conceptos semánticos de un lenguaje natural, habremos de utilizar algún lenguaje, y como el lenguaje que utilizemos para el análisis es un sublenguaje del lenguaje ordinario, el único modo de definir la semántica del lenguaje ordinario es hacerlo desde dentro de él mismo. El mentiroso reforzado se produce al intentar ser consecuente con esta idea de universalidad del lenguaje ordinario: escindiendo el lenguaje del análisis del lenguaje analizado, al modo tarskiano, la paradoja se resuelve, pero al intentar expresar en el lenguaje objeto el resultado del análisis, entonces se reproduce la paradoja. Esto no sucede en el caso de las paradojas matemáticas, pues en la teoría de conjuntos hablamos de objetos cuya representación en el lenguaje no forma parte del propio sistema de objetos: la teoría y la metateoría están ontológicamente separadas. Al efectuar una teoría de conjuntos que solucione la paradoja de Russell no se crean nuevas entidades metateóricas que deban ser incluidas en la teoría.

4) Dos tipos de problemas

Resumiendo, para Priest la paradoja de Russell y la del mentiroso tienen una unidad estructural, lo que exige una solución común, mientras que para Brendel hay una diferencia ontológica fundamental entre la teoría de conjuntos y las teorías semánticas que explica el trato diverso dispensado a ambas paradojas.

No puede negarse la igualdad estructural expuesta por Priest⁵. Sin embargo, su principio de solución uniforme se presenta más dudoso. Aun cuando ambas paradojas compartan una misma estructura, no hay ningún impedimento de principio por el que no pueda darse una definición completa de ellas interrumpiendo el argumento en pasos diferentes en cada caso: por ejemplo, negando la existencia de w en el caso de la paradoja de Russell y pasando a una lógica trivalente en la que las oraciones $\delta(V) \in V$ y $\delta(V) \notin V$ no sean ni verdaderas ni falsas, en el caso del mentiroso. Que todas las paradojas muestren un mismo tipo de razonamiento sólo permite concluir que, si se evita alguno de los pasos que conducen a él, se evitarán todas las paradojas. Pero esto no indica que la misma solución tenga sentido para todas las formas de las paradojas. De hecho, Brendel muestra que no todas las soluciones tienen los mismos criterios de adecuación.

Sin embargo, también las opiniones de Brendel deben ser matizadas. Para ello necesitamos antes establecer una distinción. Enfrentados a una paradoja, podemos plantearnos dos tipos de problemas⁶:

- El problema descriptivo, que trata de averiguar qué había de erróneo en las teorías intuitivas que conducían a las paradojas. El objetivo es crear teorías que no produzcan paradojas y que recojan las intuiciones esenciales de la teoría en cuestión.
- El problema normativo, que trata de diseñar lenguajes que estén libres de paradojas y que sean útiles para ciertos fines determinados.

Si distinguimos entre estos dos tipos de problemas, las razones ontológicas que Brendel postula para separar la paradoja de Russell de la del mentiroso dejan de ser relevantes para aceptar o no un tipo de solución, y lo que se convierte en decisivo es el fin perseguido con la solución a la paradoja. Ambas paradojas tienen una "historia" semejante. Ambas muestran que una cierta teoría intuitiva es inconsistente: la de conjuntos y la de la verdad. Ahora bien, como ya hemos comentado, la paradoja de Russell amenazaba la

⁵ Feferman (1984) estudia un lenguaje formal en el que la convención T y los axiomas de comprensión se subsumen como instancias particulares de un esquema más general, de tal modo que se muestra formalmente la unidad de las soluciones a ambas paradojas.

⁶ Véase, por ejemplo, (Gupta y Belnap, 1993), p. 10.

estabilidad de los mismos fundamentos de la matemática, por lo que el problema normativo era acuciante: era preciso encontrar una teoría consistente que pudiera demostrar la existencia de todos los conjuntos necesarios para una correcta fundamentación de las matemáticas. La solución de Russell fue renunciar a la teoría intuitiva de conjuntos y sustituirla por un sistema axiomático acompañado de una reglamentación del lenguaje, de tal modo que la oración paradójica no podía ser escrita. De un modo análogo, cuando el problema de la verdad se sitúa en el ámbito de los estudios metamatemáticos, se plantea el problema normativo de encontrar una definición de la verdad que permita construir una semántica formal para la lógica de predicados. Es necesario encontrar una definición de la verdad que sea acorde con la Convención T, pero que evite las paradojas. La solución de Tarski a este problema normativo es análoga a la de Russell, y también pasa por una regimentación del lenguaje, mediante una jerarquía de predicados verdad. Si el lenguaje contiene a su predicado verdad (es semánticamente cerrado) entonces se produce la paradoja, así que Tarski propone limitarse al estudio de lenguajes semánticamente abiertos, y definir su predicado verdad en un metalenguaje. En ambos casos, mientras el fin del estudio sea la utilidad para cierto trabajo formal, la solución de Russell (mejorada con las teorías axiomáticas de ZF o NBG) y la de Tarski no plantean problemas.

Ahora bien, si consideramos el problema descriptivo de la verdad es cuando la solución tarskiana aparece como insatisfactoria, pues la multiplicidad de predicados verdad jerarquizados no permite expresar en el lenguaje oraciones autorreferentes que son no problemáticas en el lenguaje ordinario (véase (Kripke 1975)). Además la solución está sujeta al problema del mentiroso reforzado (la oración “esta oración no es verdadera en ningún nivel de la jerarquía”). La solución no es descriptivamente adecuada porque no refleja características esenciales del funcionamiento del predicado verdad en el lenguaje ordinario. Esto, sin embargo, no es exclusivo de las paradojas semánticas debido a su carácter reflexivo, sino que sucede también en las paradojas de la teoría de conjuntos: la teoría intuitiva lleva aparejada una exigencia de universalidad que no se preserva en la solución de Russell. El principio de comprensión intuitivo exige que a toda propiedad le corresponda una extensión, es decir, el conjunto de las cosas que satisfacen esa propiedad. Formalmente, esto implica que para cada fórmula abierta del lenguaje de conjuntos exista el conjunto de las cosas que satisfacen esa fórmula. Pero esto provoca la paradoja. La diferencia entre teoría y metateoría se introduce al restringir este principio (mediante la teoría de tipos o el esquema axiomático de separación) para evitar la paradoja. Entonces no puede demostrarse la

existencia de un conjunto en la teoría para cada propiedad metateórica. La exigencia de universalidad exigiría, por ejemplo, la consideración de un conjunto de todos los conjuntos, lo que conduciría de nuevo a la paradoja.

En definitiva, los criterios de adecuación de las soluciones a las paradojas dependen de qué tipo de problema se esté intentando resolver. Para Brendel hay una razón ontológica de principio que indica que para la paradoja de Russell puede aceptarse una solución restrictiva, mientras que para la del mentiroso no. En nuestra opinión, la diferencia no reside en el tipo de concepto implicado, sino en el tipo de pretensión de la solución. Si nuestro objetivo es normativo, pueden aceptarse restricciones a la universalidad, pero si pretendemos describir las teorías intuitivas no. Podría argumentarse que en la teoría de conjuntos el interés descriptivo es nulo, y que basta con una solución jerárquica para satisfacer todos los usos interesantes de la noción de conjunto. Esto no es aceptable en el plano filosófico, pues obliga a distinguir entre las propiedades que tienen como extensión a un conjunto y aquellas que no lo tienen, con el único motivo de que estas producirían paradojas si su extensión fuese un conjunto. Ni siquiera en el plano puramente matemático es esa opinión indiscutible. Feferman (1984, pp. 237-239) ha hecho notar lo poco natural que resulta una teoría de conjuntos jerárquica como ZF o NBG para servir de fundamento a disciplinas matemáticas como la teoría de categorías, donde la autorreferencia es natural.

BIBLIOGRAFÍA

- Brendel, Elke: *Die Wahrheit über den Lügner. Eine philosophisch-logische Analyse der Antinomie des Lügners*, Walter de Gruyter, 1992.
- Feferman, Solomon: "Toward Useful Type-Free Theories, I", en R. Martin (ed.), *Recent Essays on Truth and the Liar Paradox*, Clarendon Press, 1984.
- Gupta, Anil y Belnap, Nuel: *The Revision Theory of Truth*, MIT Press, 1993.
- Kripke, Saul: "Outline of a Theory of Truth", *Journal of Philosophy*, 72, 1975, pp. 690-716.
- Priest, Graham: "The Structure of the Paradoxes of Self-Reference", *Mind*, 103, n°409, 1994, pp. 24-34.
- Ramsey, Frank P.: "The Foundation of Mathematics", en *The Foundations of Mathematics and other Logical Essays*, Braithwaite, R.B. (ed.), Kegan Paul, 1931.
- Simmons, Keith: *Universality and the Liar. An Essay on truth and the Diagonal Argument*, Cambridge University Press, 1993.

José Martínez Fernández
 Universidad de Valencia
 e-mail: martifer@uv.es

PSICOLOGISMO Y NORMATIVISMO EN LÓGICA: El Caso del Razonamiento Revocable

CARLOS A. OLLER
Universidad de Buenos Aires

ABSTRACT: Frege's critique of psychologism is sometimes considered the founding document of the analytic tradition. The philosophical debate between psychologism and normativism in logic is far from extinguished and has recently resurfaced with the advent of the logics of defeasible reasoning. In this paper I address several issues related to the nature and justification of defeasible logics and I defend a variety of the normativist stance with respect to these logics. Several arguments against the claim that, given the alleged psychologistic nature of the enterprise of formalization of defeasible reasoning, psychologism is correct with respect to defeasible logics are advanced.

KEYWORDS: Philosophy of logic, psychologism, defeasible reasoning, nonmonotonic reasoning.

1. Introducción

La crítica de Frege al psicologismo es considerada por algunos autores como el documento fundacional de la tradición analítica en filosofía. El debate filosófico entre el psicologismo y el normativismo en lógica, sin embargo, está lejos de haberse extinguido y ha recientemente resurgido con la aparición de las lógicas del razonamiento revocable (*defeasible reasoning*) o no monótono, dos tipos de razonamiento que a veces son identificados en la literatura.

El estudio formal del razonamiento revocable se originó en la segunda parte de la década del setenta en el campo de la inteligencia artificial, y ha abierto un área nueva de investigación en lógica. Los investigadores en inteligencia artificial señalaron que la lógica deductiva clásica no es suficiente para modelizar el razonamiento humano. El razonamiento humano de sentido común es con frecuencia revocable: los seres humanos extraen conclusiones racionales a partir de información incompleta en ausencia de evidencia que haría implausibles a esas inferencias.

Mientras que el razonamiento deductivo es monótono, el razonamiento revocable es una variedad de inferencia no monótona, un tipo de razonamiento caracterizado por el hecho de que la adición de nuevas premisas puede socavar conclusiones previamente obtenidas. En términos formales, una operación de consecuencia Cn definida para un lenguaje L es no monótona si no cumple con la siguiente propiedad, donde Γ y Δ son conjuntos de oraciones de L :

$$\text{Si } \Gamma \subseteq \Delta, \text{ entonces } Cn(\Gamma) \subseteq Cn(\Delta)$$

En la literatura sobre este tema, que ha alcanzado un nivel de madurez formal, se han seguido tanto tratamientos cualitativos como cuantitativos del razonamiento revocable.

En un trabajo reciente, F.J. Pelletier y R. Elio [Pelletier & Elio 97] han defendido una posición psicologista respecto al papel de los estudios empíricos sobre el razonamiento revocable humano en la creación y justificación de teorías formales sobre la inferencia revocable. Para ello, señalan que la gran mayoría de los investigadores en el área de la lógica del razonamiento revocable afirman que su motivación para formalizar este tipo de inferencia es modelizar una propiedad del razonamiento humano, su flexibilidad, que ellos consideran que está esencialmente conectada con la naturaleza revocable del razonamiento de sentido común. Pero, no sólo la motivación para construir lógicas del razonamiento revocable parece ser de naturaleza psicologista: la justificación de los patrones de inferencia de estas lógicas parece estar compelida a depender de la manera en que la gente razona revocablemente. Ésto sucede porque no parece haber un criterio de corrección para las inferencias revocables que sea al mismo tiempo universalmente aceptado por los lógicos e independiente de la manera en que la gente efectivamente razona.

En este trabajo se tratan algunos puntos relacionados con la naturaleza y justificación de las lógicas del razonamiento revocable, y defiendiendo una variedad de la posición normativista con respecto a estas lógicas. Se exponen argumentos en contra de la afirmación de que, dada la supuesta naturaleza psicologista de la empresa de formalización del razonamiento revocable, el psicologismo es esencialmente correcto respecto de estas lógicas.

2. Razones para la construcción de lógicas del razonamiento revocable

Los investigadores en el área de la inteligencia artificial señalaron que una propiedad clave de la inteligencia humana es su flexibilidad. Los seres humanos son capaces de extraer conclusiones a partir de la información de la que disponen en un momento dado, y de actuar de manera acorde; pero, también suelen retractar esas conclusiones si éstas resultan incompatibles con

la nueva información que reciben. La flexibilidad exhibida por la inteligencia humana se debe pues, según estos autores, a la naturaleza revocable del razonamiento de sentido común.

Recordemos que, según la definición de Marvin Minsky [Minsky 68], la inteligencia artificial es la ciencia de hacer que las máquinas hagan cosas que requerirían inteligencia si fueran hechas por personas. De manera que si se quiere que los programas de computadora actúen de manera “inteligente”, entonces deberán exhibir una flexibilidad similar o superior a la ostentada por los seres humanos. Dado lo expuesto en el párrafo anterior, una manera (aunque no la única concebible) de lograr esta flexibilidad pareciera ser la construcción de programas que incluyan mecanismos inferenciales para “razonar” de manera revocable. Son diversas y numerosas las áreas de posible aplicación de tales mecanismos inferenciales. Así por ejemplo, la mayor parte de las normas legales presentan excepciones no incluidas en la formulación de la norma y, por lo tanto, las lógicas del razonamiento revocable parecen ser particularmente adecuadas para la formalización del razonamiento legal.

Una vez que se ha decidido modelizar el razonamiento revocable de sentido común mediante formalismos lógicos, o basados en la lógica, es necesario plantearse cuál será la relación entre estos formalismos y los patrones inferenciales que es posible observar en los seres humanos. Pelletier y Elio muestran que un gran número de los investigadores en lógica no monótona consideran importante y necesario el estudio de la manera en que los seres humanos razonan revocablemente, no sólo para la fundamentación sino también para la justificación de las teorías formales sobre la inferencia revocable. Baste como ejemplo una cita del magistral trabajo de David Makinson sobre esquemas generales del razonamiento no monótono:

Of course, humans have been reasoning nonmonotonically for as long as they have been reasoning at all.[S]ome more or less disciplined procedures have come into use. ... Such [procedures] are not discussed explicitly in this [work], which is devoted to formal constructions for artificial intelligence. But they should not be forgotten, either, as they reveal a background of human practice with which to compare formal proposals for eventual computer implementation. [Makinson 93]

Pero, esta posición respecto de la relación entre los formalismos no monótonos y el razonamiento humano suscita nuevamente el viejo problema de la justificación de los principios lógicos, y revive la polémica entre psicologismo y normativismo en lógica.

3. Psicologismo y normativismo en la justificación de la inferencia revocable

El psicologismo respecto a un campo de conocimiento dado es el punto de vista según el cual el contenido de ese campo de conocimiento es lo que reside en los estados psicológicos de la mayor parte de los seres humanos. Así, una posición psicologista en lógica sostiene que su contenido queda definido por aquellas inferencias que la gente juzga válidas, adecuadamente generalizadas. Y, en el caso particular de la lógica no monótona, por aquellas inferencias revocables que la mayor parte de la gente acepta como correctas.

Sin embargo, pareciera que, en principio, es posible dirigir contra esta posición psicologista todos los argumentos esgrimidos en contra del psicologismo respecto de la lógica clásica [Kusch 95]. Pelletier y Elio argumentan que esto no es así porque en lógica clásica, a diferencia de lo que ocurre en las lógicas del razonamiento revocable, existe un propósito independiente de la gente que esa teoría pretende recoger: la transmisión de la verdad de premisas a conclusión. Es justamente este propósito independiente de la gente el que sirve como un ideal normativo que nos permite afirmar que la mayor parte de los seres humanos cometen errores lógicos al razonar. Por contraste, estos autores sostienen que la empresa de formalizar el razonamiento revocable es una tarea básicamente psicologista, cuyo objetivo es precisamente simular la manera en que los seres humanos razonan en un determinado tipo de situaciones. Afirman que no existe, como en el caso de la lógica clásica, un propósito independiente de la gente que permita independizar las lógicas no monótonas del razonamiento humano revocable.

Hay por lo menos dos argumentos que es posible esgrimir contra la posición de Pelletier y Elio: el primero consiste en cuestionar su premisa de que no existe para las lógicas del razonamiento revocable un criterio de corrección que sea independiente de la gente, y el segundo tiene que ver con las distinción entre lógica pura y lógica aplicada.

Una forma del primer tipo de argumento es considerado por los mismos Pelletier y Elio: es posible sostener que la inferencia humana revocable es a la teoría de la probabilidad cualitativa como el desempeño de los seres humanos respecto de la deducción es a la lógica formal deductiva. La teoría de la probabilidad cualitativa proporcionaría así un criterio independiente de la manera en la que la gente razona revocablemente que, a su vez, permitiría evaluar el desempeño humano en este campo. De este modo, las lógicas del razonamiento revocable se independizarían de los patrones inferenciales que es posible observar en los seres humanos, y la posición psicologista respecto de ellas devendría inaplicable.

Con respecto al segundo argumento, es innegable que los formalismos no monótonos han surgido como lógicas aplicadas cuya intención era modelizar el razonamiento revocable humano. Naturalmente, este origen no “condena” a la filosofía de las lógicas revocables a suscribir una posición psicologista, del mismo modo que el origen de la geometría en la agrimensura no hace a aquella una ciencia de la medición de los terrenos. En principio, es posible diferenciar la teoría pura de la aplicación que ha motivado la construcción de esa teoría; de la cual es posible separarla. Sin embargo, es preciso tener en cuenta que las lógicas revocables parecen tener con la aplicación que las ha motivado una relación especial, similar a la que se da entre la lógica deóntica y el discurso normativo. Desde un punto de vista puramente formal las lógicas deónticas habituales son sistemas intensionales que podrían estudiarse dentro del marco general de las lógicas modales. Pero, un rasgo distintivo de las lógicas deónticas es su interés en formalizar el discurso normativo y en recoger de la manera más fiel posible los fenómenos inferenciales que se dan en él. El desarrollo histórico de la lógica deóntica puede verse justamente como el resultado del movimiento entre la teoría pura y su aplicación en la formalización del discurso normativo. Algo similar sucede con las lógicas del razonamiento revocable: si bien es posible separar la teoría pura de su aplicación en la formalización del razonamiento revocable, el interés de las lógicas no monótonas parece residir en el logro esta finalidad. Pero, el interés esencial de una teoría lógica en su aplicación específica no implica automáticamente el abandono de una pretensión normativa de la teoría: si bien debe existir alguna suerte de isomorfismo entre el campo a representar y la teoría, la teoría diferir de lo representado y determinar normativamente que, en el caso que estamos tratando, ciertos esquemas inferenciales son o no son correctos. De manera que, aun aceptando que existe una relación especial entre las lógicas del razonamiento revocable y el razonamiento revocable humano, ésto no nos obliga a aceptar una posición psicologista en esta área.

BIBLIOGRAFÍA

- Kusch, M., *Psychologism*, Routledge, Londres, 1995.
- Makinson, D., “General Patterns in Nonmonotonic Reasoning”, en Gabbay, Hogger & Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol.3: *Nonmonotonic Reasoning and Uncertain Reasoning*, Oxford University Press, Oxford, 1993, pp. 35-110.
- Minsky, M. (ed.), *Semantic Information Processing*, MIT Press, Cambridge (Mass.), 1968.
- Pelletier, F.J. & Elio, R., “What Should Default Reasoning Be, By Default?”, en *Computational Intelligence*, 15:1997, pp.165-187.

Carlos A. Oller

Universidad de Buenos Aires

e-mail: coller@ciudad.com.ar



ELECCIÓN DE CONSTANTES LÓGICAS Y ELECCIÓN DE POSTULADOS SEMÁNTICOS *

MANUEL PÉREZ OTERO
Universidad de Barcelona

ABSTRACT: The standard views about the notions of logical truth and analytical truth are compared. I focus on the strong contrast between the treatments given to each of them in Quine's "Two Dogmas of Empiricism" and dispute that such a differential attitude is justified. I explain why the central Quinean criticism against analytical truth can be adapted to jeopardize the model-theoretic account of logical truth. The core idea is that in this account a non-general and non-explained language-relative enumeration of logical constants plays the same role as the usual choices of semantical postulates play in Carnap-style accounts of analytical truth.

KEYWORDS: Meaning postulate, logical constant, analyticity, language-relativity, explanation

1. Introducción

Uno de los propósitos principales de este trabajo es comparar los puntos de vista prevalecientes acerca de dos conceptos que han recibido gran atención a lo largo del siglo XX. Se trata de los conceptos de verdad lógica y verdad analítica. No sería exagerado decir que los intentos de elucidar cada una de ambas nociones han aglutinado una parte fundamental del quehacer en el ámbito de la filosofía analítica –especialmente respecto a la labor desarrollada durante los inicios y la consolidación de esa tradición– convirtiéndose en una de sus marcas distintivas más características.

Es bien sabido, sin embargo, que la opinión generalizada acerca de lo fructífero que hayan podido resultar esos esfuerzos difiere considerablemente en relación con uno u otro caso. Hay un acuerdo bastante amplio en que la explicación teórico-modelista tarskiana de la verdad lógica y la consecuencia lógica es una explicación al menos extensionalmente correcta de esos conceptos. Pero la situación es muy diferente cuando nos dirigimos a la verdad analítica. Incluso entre los filósofos que aceptan como legítima la idea de

*Este trabajo forma parte de los proyectos de investigación PB96-1091-C03-03 (subvencionado por el Ministerio de Educación y Cultura de España) y 1997SGR 00396 (subvencionado por la Generalitat de Catalunya).

analiticidad (y la noción menos técnica de significado, emparentada con ella) predomina la creencia de que nadie ha proporcionado una elucidación de ese concepto que sea comparable a la que se dispone para el concepto de verdad lógica. Eso sin contar el hecho de que otros muchos especialistas, sobre todo en el ámbito de la lógica matemática, contemplan con bastante escepticismo la posibilidad de que pueda ofrecerse un tratamiento preciso de la noción general de significado, la cual supuestamente habría de utilizarse para fundamentar la distinción entre lo analítico y lo sintético.

No es muy difícil identificar algunos de los factores más directamente responsables de tal asimetría. En la parte positiva, y como ya se ha sugerido al clasificarla, la explicación teórico-modelista de los conceptos lógicos procede fundamentalmente de Tarski y otros autores que trabajaban en su misma tradición. La labor de refinamiento técnico y conceptual llevada a cabo en el seno de esa tradición con el auxilio de la teoría matemática de conjuntos ha hecho que la teoría de modelos resultante goce del rigor y prestigio que la acreditan desde hace décadas como un cuerpo doctrinal científicamente sólido.

En la parte negativa, el fomento del escepticismo sobre las perspectivas de caracterizar la clase de las verdades analíticas (y un concepto correspondiente más general de *consecuencia* analítica) ha de ser atribuido principalmente a Quine. El argumento más fundamental de Quine contra la analiticidad es, según él mismo ha manifestado, el que se apoya en su tesis de la indeterminación de la traducción radical. Pero dicha tesis depende, como reconoce el propio Quine, de presupuestos conductistas acerca del significado lingüístico. Pocos filósofos comparten hoy ese conductismo por lo que cabe pensar que buena parte del rechazo generalizado a la analiticidad, en cuanto inspirado —más o menos explícitamente— por las críticas de Quine, ha de basarse en los otros argumentos quineanos, los cuales aparecen sintetizados en su artículo más emblemático: “Dos dogmas del empirismo”.

Para evaluar esa situación de asimetría prestaremos atención a ese artículo, resaltando el tratamiento diferencial que dispensa Quine a cada uno de los dos conceptos en cuestión: verdad lógica y verdad analítica.

2. Verdad lógica y verdad analítica caracterizadas

Antes de exponer propiamente la crítica de Quine que es aquí pertinente —cosa que haremos en la próxima sección— procederemos a describir en esbozo las líneas generales de una definición de verdad (y consecuencia) analítica. Ello permitirá situar apropiadamente dicha crítica y poder apreciar su fuerza.

La definición en cuestión depende esencialmente de otra caracterización ya mencionada: la definición teórico-modelista estándar de verdad (y consecuencia) lógica. Por esa razón, empezaremos por resumir muy sucintamente algunos puntos básicos de esta última definición.

La explicación proporcionada por el enfoque teórico-modelista tarskiano se aplica preferentemente a lenguajes formalizados en primer orden (en el sentido de que es para esa familia de lenguajes –primer orden– que la explicación resulta relativamente poco controvertida)¹. También aquí nos restringiremos a tales lenguajes. En ellos se efectúa una primera clasificación general de las expresiones básicas en dos grupos (y dejando de lado los paréntesis, que se consideran símbolos auxiliares), que integran respectivamente el vocabulario lógico y el vocabulario extra-lógico. Pertenecen al vocabulario lógico símbolos de cuatro clases: i) las conectivas sentenciales: negación, disyunción, conjunción y condicional; ii) un signo que expresa la relación de identidad; iii) los cuantificadores universal y existencial; iv) un conjunto infinito de variables individuales. Entre los símbolos extralógicos se cuentan, por una parte, los términos singulares (cuyos correlatos en los lenguajes formales se denominan ‘constantes individuales’); los otros símbolos extralógicos son los predicados, incluyendo como tales las expresiones relacionales de cualquier *ariedad* finita (la *ariedad* es el número de términos que requiere el predicado para formar una fórmula).

Verdad y consecuencia lógica se definen a partir de los conceptos semánticos de *modelo* y de *verdad en un modelo*. Dado un lenguaje L , un modelo para L es un par ordenado $\langle D, F \rangle$, en el que D , el *dominio* o universo del modelo, es un conjunto no vacío cualquiera, y F es una *función de interpretación* que asigna valores semánticos a las expresiones extralógicas de L : a cada término singular le asigna un elemento de D y a cada predicado n -ádico le asigna un conjunto de n -tuplas de elementos de A (es decir, un conjunto de secuencias de n elementos de A). Los valores de verdad de las fórmulas se relativizan a modelos (recurriendo también a *asignaciones* de valores a las variables, para el caso de fórmulas abiertas). Tras definir el concepto de verdad en un modelo (no ofrecemos aquí los detalles), se definen como verdades lógicas los enunciados que son verdaderos en todo modelo; y se caracteriza la relación de consecuencia lógica de modo que un enunciado α es consecuencia lógica de un conjunto de enunciados Σ si y sólo si α es verdadero en todos los modelos en que son verdaderos todos los enunciados de Σ .

Veamos ahora qué definición de analiticidad (verdad y consecuencia analítica) se propone aquí. La idea es muy simple. Un desiderátum obvio es que la clase de verdades lógicas quede incluida propiamente en la clase de verdades analíticas. Para ello ha de debilitarse el requisito exigido a las verdades lógicas (ser verdaderas en todo modelo), requiriendo de las verdades analíticas sólo que sean verdaderas en todos los modelos de una cierta subclase que ha de

¹ En este escueto sumario parcial prescindiremos de la última etapa en el proceso de abstracción a partir del lenguaje natural que consiste en el tránsito desde un lenguaje formalizado a un lenguaje formal.

ser especificada. La especificación se efectúa mediante lo que podemos denominar *postulados semánticos*. Cada lenguaje formalizado llevará asociado un conjunto de tales postulados. Éstos establecerán ciertas relaciones básicas entre las denotaciones de algunas expresiones extralógicas; esas relaciones son las que todo hablante del lenguaje natural (para cuya versión formalizada se ofrecen las definiciones) reconoce meramente en virtud de comprender su significado.

Un ejemplo trivial aclarará el funcionamiento de los postulados semánticos. El enunciado ‘Todo armario es un mueble’ –o más exactamente la versión formalizada del mismo: $\forall x (\text{Armario}(x) \rightarrow \text{Mueble}(x))$ – es un caso prototípico de enunciado que no es una verdad lógica pero sí es una verdad analítica. Que no sea una verdad lógica se debe (asumiendo la definición estándar) a que existen modelos cuya función de interpretación asigna a ‘Armario’ como denotación un conjunto no incluido en el conjunto asignado a ‘Mueble’. De esos modelos diremos que no son *semánticamente admisibles*, por cuanto que no respetan una relación semántica obvia, y reconocible como tal por todo usuario competente del lenguaje, entre ambos predicados. Pues bien, los postulados semánticos serán los que seleccionen, en general, cuáles habrán de ser modelos semánticamente admisibles, restringiendo así la clase más amplia de todos los modelos. Para este caso particular, por ejemplo, un postulado semántico establecería que la extensión asignada a ‘Armario’ en un modelo ha de ser un subconjunto de la extensión asignada a ‘Mueble’. Modelos cuya función de interpretación no cumpla con esa restricción no son semánticamente admisibles.

Otros postulados semánticos establecerán, por ejemplo, que ‘es mayor que’ debe denotar una relación transitiva, o quizá, aunque éste es un caso bastante dudoso, que el individuo denotado por ‘Sócrates’ ha de pertenecer al conjunto-extensión denotado por ‘Humano’.

Podemos ahora concretar más las definiciones acerca de la analiticidad. Un enunciado del lenguaje L es analíticamente verdadero (en L) si y sólo si es verdadero en todo modelo semánticamente admisible para L . La relación de consecuencia analítica se definiría también del modo que cabe esperar: un enunciado α es consecuencia analítica de un conjunto de fórmulas Σ si y sólo si α es verdadero en todos los modelos semánticamente admisibles en que son verdaderos todos los enunciados de Σ .

Ese tipo de caracterización no es original. El método se corresponde en lo sustancial con el empleado por Carnap en su breve artículo “Meaning Postulates” (Carnap, 1952). Hay algunas diferencias de detalle. La más destacable es que la explicación de la verdad lógica empleada por Carnap (pp. 223-224), y que también le sirve como primer paso para definir luego analiticidad, se ha sustituido por la explicación teórico-modelista habitual. En

parte debido a eso, es también diferente el modo de articular las restricciones semánticas impuestas por una u otra forma de postulados. Pero ninguna de esas diferencias tiene mucha importancia. En cualquier caso, no son relevantes a los efectos de nuestro presente tema. Podemos decir que disponemos de una caracterización de las verdades analíticas (y de la relación de consecuencia analítica) que, en lo fundamental, procede de Carnap. Consiste en combinar una variante de la idea carnapiana de los postulados del significado con la explicación tarskiana de la lógica que hoy es estándar (pero que no estaba suficientemente bien establecida cuando, en 1952 y también anteriormente, Carnap propugnaba su propia explicación de la verdad y la consecuencia lógica como parte de su explicación de la analiticidad). En resumen, la clase más amplia de modelos lógicamente posibles queda restringida mediante un conjunto de postulados semánticos que, al hacer explícitas ciertas relaciones entre los significados de las expresiones del vocabulario extra-lógico (aquellas relaciones que todo usuario del lenguaje, y meramente en virtud de serlo, podría reconocer) caracterizan una subclase de modelos semánticamente admisibles.

3. Explicaciones por enumeración relativa a un lenguaje

He usado ‘postulado semántico’ como etiqueta genérica relativamente neutral que pudiera servir como recordatorio no únicamente de los *meaning postulates* que propone Carnap en 1952, sino también de algunos otros de los métodos análogos que había formulado anteriormente, y que eran bien conocidos por Quine cuando éste redacta su “Dos dogmas del empirismo”. En particular, las diversas formas de *reglas semánticas* carnapianas que el propio Quine menciona (Quine 1951, sección 4) son otras tantas versiones del mismo tipo de método que representan los postulados semánticos, o los *meaning postulates*. Interesa señalar, por lo tanto, que las modificaciones que aporta Carnap en 1952 no constituyen ninguna novedad sustancial respecto a sus tesis sobre la analiticidad inmediatamente anteriores que había criticado Quine (se trata meramente de una formulación más clara de sus posiciones). En otras palabras: las diversas objeciones de Quine se aplican con la misma fortuna a las diferentes caracterizaciones de lo analítico manejadas por Carnap en esa época. Consiguientemente, su argumento crítico principal –que vamos a examinar enseguida– se aplicará del mismo modo a nuestra definición anterior.

Ahora hemos de prestar atención a tal argumento. Después de pasar revista a varias aproximaciones a la idea de analiticidad, Quine considera y rechaza, en la sección 4 –titulada “Reglas semánticas”–, las estrategias de Carnap para caracterizar los enunciados analíticos en relación con lenguajes artificiales. Vamos a resumir el núcleo de su objeción, adaptándola ocasionalmente a nuestra definición de analiticidad ya que –como hemos

indicado— si dicha objeción está bien fundamentada, entonces se aplicará igualmente contra el método de los postulados semánticos.

Quine afirma primero (como paso inicial clarificador que ayudará a presentar luego el problema) que lo que se pretende conseguir es elucidar de modo general la relación ‘ser analítico para’ que se daría entre enunciados y lenguajes; es decir, dar un sentido general a la construcción ‘*E* es analítico para *L*’, tomando ‘*E*’ y ‘*L*’ como variables (Quine 1951, p. 33).

Las reglas semánticas carnapianas de un lenguaje particular *L* (al igual que los postulados semánticos) proporcionarían un método para determinar qué enunciados de *L* son analíticos, es decir, qué enunciados son analíticos para *L*. Así pues, la definición, esquemáticamente presentada, establecería lo siguiente: un enunciado *p* es analítico para *L* si y sólo si *p* es verdadero en virtud de las reglas semánticas (o de los postulados) de *L*. En nuestro caso ‘verdadero en virtud de los postulados semánticos de *L*’ ha de entenderse como una abreviación de ‘verdadero en todo modelo seleccionado por los postulados semánticos de *L*’.

A partir de ahí se abren dos posibilidades, según cuál sea la forma de la regla semántica (o del postulado): ésta puede o no involucrar una referencia explícita al propio concepto de verdad analítica. Ni en los *meaning postulates* de Carnap ni en nuestros postulados semánticos se menciona explícitamente dicho concepto. Concentrémonos pues en la discusión y rechazo por parte de Quine de esta segunda posibilidad, en la que la regla (o postulado) se formula sin apelar a la analiticidad. Quine distingue dos modos de concebir la regla:

(i) Aunque en la formulación de la regla semántica no intervenga el concepto de analiticidad, se pretende que no cualquier verdad quede seleccionada por una de tales reglas. La regla ha de especificar sólo ciertos enunciados apropiados, los analíticos, aunque sin decirlo explícitamente. Ocurre lo mismo con los postulados semánticos, que se escogen a fin de preservar en todos los modelos semánticamente admisibles ciertas (pero no cualesquiera) relaciones que se dan entre las denotaciones de las expresiones extra-lógicas. En cualquier caso, por lo tanto, se apela a un concepto general (no relativo a un lenguaje) de regla semántica (o postulado) que, puesto que ha de satisfacer la condición recién mencionada, requiere tanta aclaración como la propia idea de verdad analítica. “Las reglas semánticas se distinguen, aparentemente, sólo por el hecho de figurar en una página bajo el rótulo ‘Reglas Semánticas’; y este rótulo carece por su parte de significado” (Quine 1951, p. 34).

(ii) Puede sostenerse, por el contrario, que el concepto de regla semántica (o postulado) sí adquiere, para cada lenguaje específico, un significado preciso. Dado un lenguaje formalizado concreto, sus reglas semánticas (o postulados) son pura y simplemente los que se declaran como tales enumerándolos explícitamente. Tenemos entonces el siguiente problema: comprendemos una

serie de conceptos concretos relativizados –‘regla semántica del lenguaje L_1 ’, ‘regla semántica del lenguaje L_2 ’, etc.– pero no tenemos comprensión sobre un concepto general de regla semántica. Se ha definido con precisión un concepto, o un conjunto de conceptos, sólo relativamente a uno o varios lenguajes específicos. Por decirlo de otro modo, cuando L_1 y L_2 son lenguajes diferentes, ninguna conexión se establece entre ‘regla semántica de L_1 ’ y ‘regla semántica de L_2 ’; ninguna aclaración acerca del concepto general ‘regla semántica de’. Eso resulta fatal teniendo en cuenta cuál era el planteamiento inicial del intento de elucidar la analiticidad: “Una vez que nos proponemos explicar ‘ E es analítico para L ’ en términos generales para la variable ‘ L ’ (incluso admitiendo la limitación de ‘ L ’ a lenguajes artificiales), la explicación ‘verdadero según las reglas semánticas de L ’ es estéril, pues el término relativo ‘regla semántica de’ necesita por lo menos tanta aclaración como ‘analítico para’” (Quine 1951, p. 34).

Puesto en otros términos, y aplicándolo al caso de nuestros postulados semánticos, que sólo adquieren significado preciso interpretándolos –conforme a (ii)– como válidos únicamente en relación con los lenguajes específicos para los cuales se eligen: ningún criterio general de carácter absoluto (en el sentido de no ser meramente relativo a un lenguaje) explícitamente declarado (y que no presuponga el propio concepto de analiticidad) guía la elección de los postulados semánticos. Por tanto, para cada lenguaje L el único sentido que tiene ‘postulado semántico de L ’ es el de servir de etiqueta para coleccionar bajo ella una serie de restricciones sobre la clase de las verdades lógicas de L al efecto de obtener una subclase de enunciados a los que se les denomina ‘verdades analíticas de L ’.

En definitiva, nuestra explicación de la analiticidad no tiene la adecuada generalidad que intuitivamente creemos reconocer en el concepto preteórico supuestamente explicado. Se trata de una explicación en la cual, ante un lenguaje formalizado determinado L , se declaran como analíticos algunos de sus enunciados por el efecto conjunto de las definiciones teórico-modelistas estándar y la elección por enumeración de ciertos postulados para L –denominados ‘semánticos’– que restringen las denotaciones del vocabulario extra-lógico permitiendo caracterizar una subclase de modelos, a los que llamamos ‘semánticamente admisibles (para L)’.

4. Logicidad y analiticidad

Como seguramente habrá quedado claro, asumo que la caracterización carnapiana de las verdades analíticas de un lenguaje que aquí se ha presentado está sujeta a esa objeción fundamental de Quine². Mi propósito no es

² Puede pensarse que, incluso restringiéndonos a “Dos dogmas del empirismo”, hay al menos otra objeción quineana no menos importante, constituida por el rechazo a la noción de verdad

salvaguardarla de tal crítica, sino sugerir que la crítica puede adaptarse para ser aplicada con la misma fuerza contra un estadio anterior en la construcción de dicha definición, en particular contra la explicación teórico-modelista de la verdad (y la consecuencia) lógica. Si esa sugerencia hipotética es acertada ello significará que, por lo que concierne al problema que aquí se ha planteado, el concepto de verdad lógica no está en mejor situación que el de verdad analítica.

Enunciada muy concisamente, la consideración que apoya dicha hipótesis es la siguiente. La explicación teórico-modelista de las nociones de verdad y consecuencia lógica para un lenguaje determinado *L* depende de una especificación mediante enumeración de cuáles son las expresiones de *L* que se consideran pertenecientes al vocabulario lógico, a las que se denomina 'constantes lógicas'³. La elección de tales constantes lógicas es relativa al lenguaje de que se trate; y ni dicha elección ni la atribución de significados particulares a cada una de esas constantes están explícitamente guiadas por un criterio general (trans-lingüístico) no relativizado más de lo que pueda estarlo la elección de nuestros postulados semánticos. En definitiva, también en la caracterización estándar de las verdades lógicas se apela a una selección por enumeración de ciertos ítems en los que reconocemos intuitivamente una determinada propiedad, para la cual no disponemos de ninguna explicación general.

Que la definición estándar de verdad lógica depende de ese modo de una adecuada especificación, relativa a un lenguaje, de las constantes lógicas queda algo oscurecido por la manera en que es presentada habitualmente dicha definición. Para empezar, la doble abstracción a partir de un lenguaje natural que supone concentrarse en una versión formalizada del mismo primero e incluso finalmente en un lenguaje formal (en que letras esquemáticas ocupan el lugar de las expresiones extra-lógicas) conlleva dar por sentado una cierta *forma* de los enunciados, lo cual ya nos compromete con la elección de un vocabulario lógico integrado por las expresiones que se mantienen constantes (a fin de preservar la forma lógica del enunciado), en contraste con las expresiones extra-lógicas (que corresponden —siguiendo la metáfora forma/contenido— al *contenido*, cambiante a través de las reinterpretaciones).

Además, el significado de tales constantes no queda explicitado asignándoles un valor semántico apropiado mediante la función de

a priori, que sustenta el *segundo dogma*. Es cierto, pero eso afecta a la analiticidad sólo debido al programa carnapiano empirista de reducción de lo *a priori* a lo analítico. Las propuestas de Carnap para caracterizar la analiticidad interesan al margen de su adhesión a ese empirismo radical; después de todo, un claro predecesor de ese proyecto carnapiano (de promoción y utilización de la idea de analiticidad como verdad en virtud de significado) es Frege, que no abraza dicha forma de empirismo.

³ Dejamos de lado las variables que, aunque se consideran pertenecientes al vocabulario lógico, ciertamente no son constantes lógicas.

interpretación de cada modelo, como sucede con las otras expresiones, sino que ese significado se pone de manifiesto (podría decirse que se “muestra”) en las cláusulas recursivas de la definición del concepto técnico de verdad en un modelo.

Aunque la concepción de la lógica que sostiene Quine en “Dos dogmas del empirismo” no coincide con la explicación teórico-modelista, lo cierto es que ambas comparten el rasgo que acabo de describir y que las exponen abiertamente a la misma crítica que Quine dirige contra la noción de verdad analítica (la crítica que hemos resumido). Efectivamente, en ese artículo Quine, lejos de intentar caracterizar de modo general (no relativo a un lenguaje) las nociones lógicas, escribe: “si suponemos un inventario previo de partículas lógicas”, y a continuación se adhiere a una versión lingüística de la concepción sustitucional originada en Bolzano (que es, a su vez, un claro precedente de la concepción teórico-modelista): “una verdad lógica es un enunciado que es verdadero y sigue siéndolo para cualquier reinterpretación de sus componentes que no sean partículas lógicas” (pp. 22-23).

Así pues, por lo que respecta a “Dos dogmas del empirismo”, mi tesis central es que cuando Quine adopta esa actitud diferencial hacia la lógica y la analiticidad lo hace sin estar respaldado por alguna razón que lo justifique. Dicho escuetamente, Quine no ofrece ninguna explicación de por qué, ante un lenguaje determinado, podemos dar por supuesto “un inventario previo de partículas lógicas” pero no un inventario previo de reglas o postulados semánticos.

La tesis puede generalizarse. Si nuestra aceptación de la explicación teórico-modelista no está respaldada por alguna propuesta sobre qué características distinguen a las constantes lógicas del resto de las expresiones, entonces estamos expuestos a un argumento análogo al de Quine, un argumento que impugnaría esa explicación por no ser suficientemente general. La reflexión puede llevarnos a concluir que ese género de argumento es correcto en lo fundamental. Y aceptar esa conclusión es compatible con la idea de que podemos darnos por relativamente satisfechos por ese tipo de método consistente en caracterizar el conjunto de constantes lógicas de un lenguaje dado (y por esa vía caracterizar ulteriormente el conjunto de sus verdades lógicas y su relación de consecuencia lógica) mediante la mera enumeración de las mismas. A pesar de sus limitaciones, dicho método –mantendría el defensor de esta idea– sería considerablemente iluminador acerca de la verdad y la consecuencia lógica.

La pregunta que se plantea aquí es la siguiente: ¿por qué no podríamos mantener la misma posición cuando se trata de la enumeración de los postulados o reglas semánticas de un lenguaje dado (que permitirán caracterizar ulteriormente el conjunto de sus verdades analíticas y su relación de consecuencia analítica)? No parece que sea defendible un tratamiento

diferenciado en lo que concierne a esta cuestión, ante la cual verdad lógica y verdad analítica no están en situación significativamente distinta⁴.

BIBLIOGRAFÍA

- Carnap, Rudolf: "Meaning Postulates", en R. Carnap, *Meaning and Necessity*, Chicago, The University of Chicago Press, 1956, 2ª edición ampliada, pp. 222-229 (publicado originalmente en 1952).
- Quine, Willard V.: "Two Dogmas of Empiricism", en W. V. Quine, *From a Logical Point of View*, Cambridge, Harvard University Press, 1953, pp. 20-46 (publicado originalmente en 1951).
- Quine, Willard V.: *Philosophy of Logic*, New Jersey, Prentice Hall, 1970.
- Tarski, Alfred: "What is a Logical Notion", *History and Philosophy of Logic* 7, 1986, pp. 143-154.

Manuel Pérez Otero
Universidad de Barcelona.
e-mail: perez@trivium.gh.es

⁴ Naturalmente existen intentos de caracterizar de modo general la logicidad (el rasgo distintivo de las constantes —o las nociones— lógicas). Uno bien conocido fue sugerido por Tarski (Tarski 1986); y el propio Quine ha hecho propuestas en esa dirección (por ejemplo, en Quine 1970, cap. 2). En la medida en que alguno de esos enfoques fuera reconocido como apropiado y se careciera de una aproximación análogamente prometedora al tema del significado y la analiticidad, la asimetría que aquí se pone en cuestión estaría justificada.

DUAL INTUITIONISTIC PARAConsISTENCY WITHOUT ONTOLOGICAL COMMITMENTS

SHAHID RAHMAN

Universität des Saarlandes (Saarbrücken, Germany) / Université Nancy 2 (Nancy, France)

JORGE ALFREDO ROETTI

Universidad Nacional del Sur (Bahía Blanca, Argentina)

ABSTRACT: The issue of this paper is to offer two (dialogical) ways of defending a non-committal interpretation of paraconsistency: we will call one the *permissive* interpretation and the other the *non-ontological commitment* interpretation. Stating contradictions and negations is from a *permissive* point of view of paraconsistency a purely formal matter: If you do so, so can I.

The *non-ontological commitment* approach results from two rules. One restricts the use and introduction of singular terms to its formal use. The other establishes how to combine the permissive interpretation with this restriction in a second order free logic.

KEYWORDS: Logic, contradiction, inconsistency, paraconsistency, dialogical logic, free logic.

Introduction

More than fifty years ago an unusual challenge appeared in analytic philosophy: In a now famous meeting of the Mathematical and Natural Science Department of the University of Torun on Friday 19 March 1949 the Polish logician Stanisław Jaskowski presented a paper entitled *A propositional Calculus for Inconsistent Deductive Systems*, which contested the Principle of Non-Contradiction. The Brazilian logician Newton C. A. da Costa introduced paraconsistent logic in his PhD in the 60s (da Costa [1974], in English) and since then developed many paraconsistent systems. Nowadays two main semantical approaches can be distinguished. One, which we call the *compelling interpretation*, based on a naive correspondence theory, stresses that paraconsistent theories are ontologically committed to inconsistent objects. The other, which we call the *non-committal interpretation*, does not assume this ontological commitment. The issue of this paper is to offer two

(dialogical) ways of defending a non-committal interpretation of paraconsistency: we will call one the *permissive* interpretation and the other the *non-ontological commitment* interpretation. Stating contradictions and negations is from a permissive point of view of paraconsistency a purely formal matter: If you do so, so can I.

The non-ontological commitment approach results from two rules. One restricts the use and introduction of singular terms to its formal use yielding a type of free logic.. The other establishes how to combine the permissive interpretation with this restriction on singular terms.

1 The permissive interpretation of paraconsistency

1.1 Dialogical logic

Dialogical logic, suggested by Paul Lorenzen in 1958 and developed by Kuno Lorenz in several papers from 1961 onwards¹, was introduced as a pragmatic semantics for both classical and intuitionistic logic.

The dialogical approach studies logic as an inherently pragmatic notion using an overtly externalised argumentation formulated as a *dialogue* between two parties taking up the roles of an *Opponent* (**O** in the following) and a *Proponent* (**P**) of the issue at stake, called the principal *thesis* of the dialogue. **P** has to try to defend the thesis against all possible allowed criticism (*attacks*) by **O**, thereby being allowed to use statements that **O** may have made at the outset of the dialogue. The thesis *A* is logically valid if and only if **P** can succeed in defending *A* against all possible allowed criticism by **O**. In the jargon of game theory: **P** has a *winning strategy* for *A*. We will now describe an intuitionistic and a classical dialogical logic, starting with the intuitionistic version.

Suppose the elements of first-order language are given with small letters (*a, b, c, ...*) for elementary formulae, capital italic letters for formulae that might be complex (*A, B, C, ...*), capital italic bold letters (***A, B, C, ...***) for predicates and τ_i for constants. A dialogue is a sequence of formulae of this first-order language that are stated by either **P** or **O**². Every move - with the exception of the first move through which the Proponent states the thesis - is an aggressive or a defensive act. In dialogical logic the meaning in use of the logical particles is given by two types of rules which determine their *local* (*particle rules*) and their *global* meaning (*structural rules*).

The particle rules specify for each particle a pair of moves consisting of an attack and (if possible) the corresponding defence. Each such pair is called a

¹ Lorenzen/Lorenz [1978].

² Sometimes we use **X** and **Y** to denote **P** and **O** with **X** ≠ **Y**.

round. A round is *opened* by an attack and is *closed* by a defence if one is possible.

$\neg, \wedge, \vee, \rightarrow; \wedge, \vee$	ATTACK	DEFENCE
$\neg A$	A	$\otimes$ (The symbol ' $\otimes$ ' indicates that no defence, but only counterattack is allowed)
$A \wedge B$	$?L(ef)$ ----- $?R(ight)$ (The attacker chooses)	A ----- B
$A \vee B$	$?$	A ----- $!_m B$ (The defender chooses)
$A \rightarrow B$	$?A$	$!_m B$
$\wedge_x A$	$?_\tau$ (The attacker chooses)	$A[\tau/x]$
$\vee_x A$	$?$	$A[\tau/x]$ (The defender chooses)

Fig. 1: Particle Rules

The first column in Fig. 1 contains the form of the formula in question, the second one possible attacks against this formula, and the last one possible defences against those attacks. (The symbol " $\otimes$ " indicates that no defence is possible.) Note that for example " $?L$ " is a move - more precisely it is an attack - but not a formula. Thus if one partner in the dialogue states a conjunction, the other may initiate the attack by asking either for the left side of the conjunction ("show me that the left-hand side of the conjunction holds", or " $?L$ " for short) or the right-hand side ("show me that the right side of the conjunction holds", or " $?R$ "). If, on the other hand, one partner in the dialogue states a disjunction, the other may initiate the attack by asking to be shown any side of the disjunction (" $?$ ").

Next, we fix the way formulae are sequenced to form dialogues with a set of structural rules (orig. *Rahmenregeln*):

R0: Formulae are alternately uttered by **P** and **O**. The *initial formula* is uttered by **P**. It provides the topic of argument. Every formula below the initial formula is either an attack or a defence against an earlier formula stated by the other player.

R1: P may only repeat attacks if the situation has changed: a situation is *changed* (allowing the repetition of an attack) if and only if O has introduced a new atomic formula (which can now be used by P).³ (No other repetitions are allowed.)

R2 (formal rule for atomic formulae): P may not introduce atomic formulae: any atomic formula must be stated by O first.

R3 (winning rule): X wins iff it is Y's turn but he cannot move (whether attack or defence).

R4 (intuitionistic rule): In any move, each player may attack a (complex) formula asserted by his partner or he may defend himself against *the last not already defended* attack. Only the latest open attack may be answered: If it is X's turn at position n and there are two open attacks m, l such that $m < l < n$, then X may not defend against m⁴.

Fig. 2: Structural rules

These rules define an intuitionistic logic. To obtain the classical version simply replace R₄ by the following rule:

R_{c4} (classical rule): In any move, each player may attack a (complex) formula asserted by his partner or he may defend himself against *any attack* (including those which have already been defended).

Fig. 3: Classical structural rule

As already mentioned, validity is defined in dialogical logic via winning strategies of P:

Def. validity: In a certain dialogical system a formula is said to be valid iff P has a (formal) winning strategy for it, i.e. P can in accordance with the appropriate rules succeed in defending *A* against all possible allowed criticism by O⁵.

1.2 The dialogical approach to permissive paraconsistency

An important distinction in dialogical logic which brings us to paraconsistency is the difference between a *provable* and a *refutable* thesis. Clearly if the Proponent has a (formal) winning strategy, the thesis is provable. But if the Opponent has a (formal) winning strategy after changing formal rights and duties, i.e. if the Opponent is the one who may use an atomic statement if and only if the Proponent has already stated the same statement before, we say that the thesis is *refutable*. Thus, clearly, contradictions are refutable. The question is whether some changes can be

³ For a precise definition of *change of situation* cf. Rahman/Rückert [1999].

⁴ Notice that this does not mean that the last open attack was the last move.

⁵ Actually we only present the *symmetric* versions of these rules. In this formulation the rules for complex propositions do not distinguish between Proponent and Opponent moves. Lorenzen and Roetti following Lorenzen usually prefer a so-called *strict* version of the *asymmetric* formulation which introduces a distinction between the Opponent and the Proponent: Both of the argumentation partners may only defend the last attacked formula but the Proponent may also attack a formula which is not the last one (cf. Roetti 1997, p. 64.) It can be shown that in intuitionistic logic the symmetric rules produce the same theorems as the asymmetric (cf. Barth/Krabbe [1982], Krabbe [1985], Felscher [1985] and Rahman [1993]).

introduced in the structural rules so as to make contradictions neither provable nor refutable: that is, $\neg\wedge A(A\wedge\neg A)$ should hold instead of $\wedge A\neg(A\wedge\neg A)$ ⁶. Such changes can indeed be introduced and they yield a dialogical formulation of paraconsistency. Rahman/Carnielli [1998] introduced a dialogical formulation of the permissive interpretation of paraconsistency with the help of the following rule:

Formal rule for negative literals: The Proponent is allowed to attack the negation of an atomic (propositional) statement if and only if the Opponent has already attacked the same statement before.

This rule yields a permissive interpretation of paraconsistency in the sense that if the Opponent concedes a contradiction, the Proponent can state it too: Stating contradictions and negations is a purely formal matter: If you do so, so can I. This idea was also expressed in Roetti [1997] in a slightly different way: if someone attacks a negation he thereby concedes that the principle of non-contradiction holds for this particular proposition. That is why if the Opponent states a contradiction based on this literal the Proponent is then allowed to attack it⁷ - Roetti speaks then of *singular paraconsistency*⁸. Notice that this rule yields a paraconsistent logic for literals only. To extend paraconsistency to complex contradictions the following rule has to be added:

Dual intuitionistic rule: In any move, each player may attack *only the last* (complex) formula asserted by his partner or he may defend himself against any attack.

This rule was stated for the first time by Lorenz in his dissertation of 1961. Lorenz calls it the *antiefektive* (i.e. antiintuitionistic) rule, but he did not seem to realise then the connections of the resulting system with Jaskowski's and da Costa's work on logics allowing inconsistencies. In the previously mentioned paper of 1997 Roetti called it the *dual intuitionistic* rule. This denomination was introduced, as far we know, in 1981 by N. D. Goodman for a non-dialogical paraconsistent system.

We formulate now this rule more precisely:

Dual intuitionistic rule:

- In any move, each player may defend himself against any attack or he may attack *the last* (complex) defence.

⁶ Cf. Roetti [1997]. Actually this leads to the standard definition of paraconsistent logics which distinguishes inconsistency of triviality (cf. da Costa / Bueno / French [1998a], 46).

⁷ These ideas can also be related with da Costa's and Arruda's definition of the paraconsistent negation. $\neg^*A =_{\text{def.}} \neg A \wedge \neg(\neg A \wedge A)$ (cf. da Costa [1980], 241 and Arruda [1984], 14).

⁸ Roetti speaks also of a *universal* paraconsistency. The idea behind this type of paraconsistency can be expressed as follows: If one argumentation partner attacks a negation he thereby concedes that the principle of non-contradiction holds for *every* formula. This yields a dialogical system where the Proponent is allowed to attack a negation iff the Opponent has attacked any negation before. The corresponding system can be formulated as a minimal logic.

- After an attack of the Proponent on move n of the Opponent no other formula stated by the Opponent before n may be used.

The result of combining the literal with the dual intuitionistic rule yields a new dialogical system of paraconsistent logic⁹. Let us consider two examples:

<i>Opponent</i>		<i>Proponent</i>	
(1) $a \wedge \neg a$	0	$(a \wedge \neg a) \rightarrow \neg a$	(0)
(3) a	2	$\neg a$	(2)
(5) $\neg a$		$\otimes$	
$\otimes$		1 ?R	(4)
		5 a	(6)
		<i>The Proponent wins</i>	

Example 1

Notation: The numbers between brackets keep track of the moves. The numbers without brackets indicate which move of the partner is thereby attacked (defences have no such numbers).

The Proponent wins because move 3 allows him to start an attack on move 6.

<i>Opponent</i>		<i>Proponent</i>	
(1) a	0	$a \rightarrow (\neg(b \rightarrow a) \rightarrow c)$	(0)
(3) $\neg(b \rightarrow a)$	2	$\neg(b \rightarrow a) \rightarrow c$	(2)
$\otimes$		3 $(b \rightarrow a)$	(4)
(5) b	4		
		<i>The Opponent wins</i>	

Example 2

The Opponent wins because the Proponent is not allowed to use a .

This paraconsistent logic actually combines two different ways of dealing with contradictions. One way - namely the literal rule - blocks or isolates the contradiction so that it cannot produce any harm. The other allows retractions - the dual rule. Because you may retract a proposition stated early the attacker is not allowed to recall the other proposition stated before and now retracted. In fact, complex contradictions rather than literal contradictions

⁹ Strategies systems for these logics are easily done by adapting the tableaux of Rahman/Carnielli [1998] to the present combination of logics. . Actually we present this system to keep the argumentation line simpler. In fact this combination of rules does not quite do the job of extending the effects of the literal rule to the complex case. The problem can be exemplified with the formula $\neg A \rightarrow (A \rightarrow b)$ - where A is a conjunction of two atomic formulae. Here the Proponent wins although - if we consider the literal case - it should not. One way to avoid this is adding the following rule: Only the last open round may be defended by the Proponent. Notice that this only applies to opens rounds, nevertheless defences of closed rounds can be repeated (of course, only if the repetition yields a change of the dialogical situation).

are in this view related to retractions and related actions of reviewing argumentative moves.

Another way to see the literal rule is to think of it as distinguishing between the internal or copulative negation from the external or sentential negation.¹⁰ That is, in the standard approaches to logic, the elementary proposition A_n has the internal logical form: $n\epsilon A$ (where ϵ stands for the copula: n *is* A) and the negation of it the form: $n\epsilon' A$ (n *is not* A). Now in this standard interpretation the negative copula is equivalent to the expression $\neg A$, where A can also be complex. Such an equivalence rejects the distinction between the internal (copulative) form and the external or sentential form of elementary propositions. This seems to be the core of an interpretation of paraconsistency which takes seriously da Costa's proposal of paraconsistent logic free of ontological commitments¹¹. Now it makes sense to ask for the ontological commitment (or non-commitment) of elementary contradictions not of complex ones. The distinction between a copulative and a sentential negation reflects this fact.

2 Paraconsistency without ontological commitments

The issue of this chapter is a dialogical formulation of free logic which allows a straightforward combination of paraconsistent and intuitionistic logic - this combination was called by Rahman *Frege's Nightmare*¹². The ideas behind this combination can be expressed in a few words: in an argumentation, it sometimes makes sense to restrict the use and introduction of singular terms in the context of quantification to a formal use of those terms. That is, the Proponent is allowed to use a constant for a defence (of an existential quantifier) or an attack (on a universal quantifier) iff this constant has been already introduced by the Opponent's attack (on a universal quantifier) or the Opponent's defence (of an existential quantifier). When the Opponent concedes any constant occurring in an atomic formula he concedes *tertium non datur* (for this formula) too. This yields a free logic which combines classical

¹⁰ The difference between internal and external negation has been worked out for other purposes by A. A. Sinowjew (Sinowjew [1970] and Wessels/Sinowjew [1975]).

¹¹ I suggest to address the inconsistency issue differently. We may well explore the rich representational devices allowed by the use of paraconsistency in inconsistent domains, but withholding any claim to the effect that there are 'inconsistent objects' in reality. (Da Costa [1998], 33).

This allows for the accommodation of inconsistency by acknowledging that it is not a **permanent** feature of **reality** to which theories must correspond, but is rather a **temporary** aspect of such **theories** [...]. In this view, to accept a theory is to be committed, not to believing it to be true *per se*, but to holding it as *if* it were true, for the purposes of further elaboration, development and investigation. (Da Costa / Bueno / French [1998b], 616-617).

¹² Cf. Rahman [1999b].

logic (for propositions with singular terms for realities) with intuitionistic logic (for propositions with singular terms for fictions)¹³. The idea now is to combine the concept of the formal use of constants in free logics and that of the formal use of elementary negations in paraconsistent logics.

Thus, the non-ontological commitment approach results from two rules. One restricts the use and introduction of singular terms to its formal use, the other establishes how to combine the literal rule with this restriction.

Formal use of constants: The Proponent is allowed to use a constant for a defence (of an existential quantifier) or an attack (on a universal quantifier) iff this constant has been already introduced by the Opponent's attack (on a universal quantifier) or the Opponent's defence (of an existential quantifier).

Non-ontological commitment rule: The literal rule applies only to formulae in which constants occur that have not been introduced in the sense of the rule for the formal use of constants.

These rules have the awkward effect that we have to decide between an intuitionistic and a classical version of the resulting logic. Now as Stephen Read remarked in his book *Thinking about Logic* the realist position in classical logic has some plausibility for existents which it definitively lacks for non-existents, fictional and beyond. Consider the proposition *Don Quijote could whistle very loudly*. Is this proposition the case or not? If one seriously believes Don Quijote is a fictional creature, one must prepared for there being no answer to this question¹⁴. Why should we presuppose tertium non datur for non-existents? One possible way of handling this has been mentioned already: the classical rule applies to existents and the non-classical to non-existents. But how to implement this in the dialogical formulation of paraconsistent free logic? In answering this question we will follow an idea of the Spanish philosopher Francisco Suárez (1548-1647) about the non applicability of the tertium non datur to propositions containing *privative* predicators. That is, Suárez distinguishes - as Aristoteles did before - predicators which can naturally be said or not of an object from those that can not be naturally applied to this object. The predicator *blind* for example can be said of Oedipus but not of a stone: You can say of a man that he is blind because the contrary is normally the case. Suárez followed from this understanding of privative predicators that tertium non datur does not apply in those propositions containing such predicators. Thus, some predicators apply and it can be decided if a given proposition built up from these predicators is the case or not. But some other predicators do not apply, for example it does not make any sense to say of Oedipus that he is an even

¹³ Cf. Rahman/Rückert/Fischmann [1999] and Rahman [1999a].

¹⁴ Cf. Read [1994], 137.

number or not. In other words, if you are allowed to use arbitrary predicates tertium non datur applies, if not (as in the case of propositions with privative predicates) then it does not¹⁵. Suárez used his theory of *privaciones* for existents. We do not. Nevertheless we will take his advice and will not presuppose in the context of dialogical free logic that any predicate can be applied to any object. We will presuppose only that any predicate can be applied to any existent object. In the language of dialogical logic: if the Opponent conceded the existence of a given object by introducing a singular term he also conceded that any predicate or its negation can be said of this object.

One way to implement this is to add to the dialogical intuitionistic paraconsistent free logic the following structural rule:

Tertium non datur for constants which have already been introduced: Once the Opponent has introduced a constant he has also conceded that tertium non-datur holds for any predicate for this constant (we express this concession for an introduced constant τ with the disjunction $\mathfrak{R}_\tau \vee \neg \mathfrak{R}_\tau$ where $\mathfrak{R}$ works as a variable for a predicate which the Proponent can substitute for any suitable predicate)

This rule allows the Proponent to choose any predicate at his convenience for the attack on the conceded tertium non-datur. This can be made clear with an example (recall that we are playing with the intuitionistic rule):

Opponent			Proponent	
(1)	$?_\tau[\mathfrak{R}_\tau \vee \neg \mathfrak{R}_\tau]$	0	$\wedge_x(A_x \vee \neg A_x)$	(0)
(3)	?	2	$A_\tau \vee \neg A_\tau$	(2)
(5)	A_τ		A_τ	(6)
			1 $?_{1/A}$	(4)
			The Proponent wins	

Example 3

In move 1 the Opponent introduced the constant τ and conceded with this introduction the disjunction $\mathfrak{R}_\tau \vee \neg \mathfrak{R}_\tau$. In move 4 the Proponent attacks this disjunction choosing the predicate A as substitution for the predicate variable $\mathfrak{R}$ and wins - the reader can easily verify that if in move 5 the Opponent chooses to defend $\neg A_\tau$ he loses too.

One important aim of the present paper and other related papers which are due to appear in a special issue of *Synthese* under the title *New Perspectives in Dialogical Logic*¹⁶ is to show how to build with the help of dialogical logic a common semantic language for different non-standard logics in such a way that 1. the semantic intuitions behind these logics can be made

¹⁵ Suárez [1960], Disputatio LIV, Sectio IV, VII, 432-438.

¹⁶ Cf. Rahman/Rückert [2000].

transparent, 2. combinations between these logics can be easily achieved, 3. a common basis is proposed for discussion of the philosophical consequences of these logics - the philosophical point here being to undertake the task of discussing the semantics of non-classical logics from a pragmatical point of view which commits itself neither to a correspondence theory of truth nor to a possible-world-semantics. But here we arrived at the end of this paper, which we hope will lead to a fruitful discussion tolerant enough to admit some but not all sorts of contradictory arguments.

REFERENCES

- Arruda, Ayda [1984]: "Aspects of the Historical Development of Paraconsistent Logic". In Priest, Routley, Norman [1984] (eds.): *Paraconsistent Logic*, München, Philosophia Verlag, 1984.
- Da Costa, N. C. A. [1974]: "On the theory of inconsistent formal systems". *Notre Dame Journal of Formal Logic*, 15, 497-510.
- [1980]: *Ensaio sobre os fundamentos da lógica*, São Paulo, Hucitec.
- [1998]: "Paraconsistent Logic". In: *Stanislaw Jaskowski Memorial Symposium. Paraconsistent Logic, Logical Philosophy, Mathematics & Informatics at Toruń*, 29-35.
- Barth, E. M. / Krabbe, E. C. W. [1982]: *From Axiom to Dialogue: A Philosophical Study of Logics and Argumentation*. Berlin/N. York: De Gruyter.
- Da Costa, N. C. A. / Bueno, O. / French, S. [1998a]: "Is there a Zande Logic?" *History and Philosophy of Logic*, XIX n. 1, 41-54.
- [1998b]: "The Logic of Pragmatic Truth". *Journal of Philosophical Logic*. 603-620, XXVII.
- Felscher, W. [1985]: "Dialogues, strategies and intuitionistic provability". *Annals of Pure and Applied Logic*, vol. 28, 217-254.
- Goodman, N.D. [1981]: "The logic of contradiction". *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 27, 119-126.
- Jaskowski, S. [1948]: "Rachunek zdan' dla systemów dedukcyjnych sprzecznych". *Studia Soc. Scient. Torunensis*, Sec. A 1, n.5, 55-77 (English translation in [1969]).
- [1969]: "Propositional Calculus for Contradictory Deductive Systems" (english translation of [1948]). *Studia Logica*, 24, 1969, 143-157.
- Krabbe, E. C. W. [1985]: "Formal Systems of Dialogue Rules". *Synthese*, N0. 63 (3), 295-328.
- Lorenz, K. [1961]: *Arithmetik und Logik als Spiele*. Diss. U. Kiel.
- Lorenzen, P. / Lorenz, K. [1978]: *Dialogische Logik*. Darmstadt: Wissenschaftliche Buchgesellschaft
- Rahman, S. [1993]: *Über Dialoge, protologische Kategorien und andere Seltenheiten*. Frankfurt a. M. / Berlin / New York / Paris / Wien: Peter Lang.
- [1999a]: "Ways of understanding Hugh MacColl's concept of symbolic existence". Appears this year in the *Nordic Journal of Philosophical Logic*.
- [1999b]: "On Frege's Nightmare". Appears in H. Wansing [2000]: *Essays on non-classical logic*. Oxford: Oxford University Press.
- Rahman, S. / Carnielli, W.A. [1998]: "The Dialogical Approach to Paraconsistency". *FR 5.1 Philosophie, Universität des Saarlandes*, Bericht Nr. 8, July 1998 (appears also in: D. Krause [2000]: *The work of Newton da Costa*, Rio de Janeiro: Kluwer).
- Rahman, S / Rückert, H. [1999]: "Dialogische Modallogik (für T, B, S4 und S5)". Appears in *Logique et Analyse*.

- [2000]: *New Perspectives in Dialogical Logic*. To appear as special issue of *Synthese* with contributions of P. Blackburn, D. Gabbay/J. Woods, J. Hintikka, E. Krabbe, K. Lorenz, U. Nortmann, H. Prakken, S. Rahman/H. Rückert, N. Rao and G. Sandu.
- Rahman, S. / Rückert, H. / Fischmann, M. [1999]: "On Dialogues and Ontology. The Dialogical Approach to Free Logic". Appears this year in *Logique et Analyse*.
- Read, S. [1994]: *Thinking About Logic*. Oxford / New York: Oxford University Press.
- Sinowjew A. A. [1970]: *Komplexe Logik. Grundlagen einer logischen Theorie des Wissens*. Berlin / Braunschweig / Basel: VEB Deutscher Verlag der Wissenschaften / Friedr. Vieweg + Sohn GmbH / Verlag und C. F. Winter'sche Verlagshandlung.
- Sinowjew, A. A. / Wessel H. [1975]: *Logische Sprachregeln. Eine Einführung in die Logik*. München / Salzburg: Wilhelm Fink Verlag.
- Roetti, Jorge A. [1997]: "Der Satz vom Widerspruch: dialogische und pragmatische Begründung". In N. Offenberger and A. Vigo [1978]: *Südamerikanische Beiträge zur modernen Deutung der Aristotelischen Logik*, Hildesheim, Georg Olms Verlag, 49-81.
- Suárez, F. [1960]: *Disputationes Methaphysicae - Disputaciones Metafisicas*. Madrid: Gredos - Biblioteca Hispánica de Filosofía.

Shahid Rahman
Universität des Saarlandes
e-mail: s.rahman@rz.uni-sb.de

Jorge Alfredo Roetti
Universidad Nacional del Sur
e-mail: usroetti@criba.edu.ar



FINITUD, VERDAD Y VAGUEDAD

FRANCISCO SALTO
Universidad de Salamanca

ABSTRACT: Finitary simulations of infinite mathematical objects are studied and applied to the understanding of sorites phenomena. Alternative set theoretical models of Mycielski's axioms in finite mathematics are studied and applied to achieve a purportedly direct, coherent and natural approach to sorites paradox. A general conceptual framework for them is presented and discussed. Finally, it is shown how to avoid sorites conditionals in these non standard finitary settings.

KEYWORDS: Sorites paradox, strict finitism, alternative set theory

0

Nociones como 'verdad' y 'vaguedad' son tan amplias y controvertidas que encontrar vínculos entre ambas resulta filosóficamente trivial. Ahora bien, otra cosa es la relación entre las paradojas lógicas o antinomias que ambas nociones pueden generar: las (familias de) paradojas del mentiroso y las (familias de) paradojas sorites. El solo hecho que ambas nociones puedan generar inconsistencias clásicas indica, por supuesto, sólo una relación superflua entre ambas. Entiendo que un índice fiable de relaciones substantivas entre ambas paradojas es la posibilidad de que ambas reciban un tratamiento *unificado*. Problemas que admiten una misma estrategia de solución son, seguramente, versiones de un mismo problema más simple y abstracto (al menos cuando éste es un problema de carácter lógico).

Las estrategias supervaluacionales han sido los instrumentos más exitosos de los que han dispuesto lógicos filosóficamente sensibles para salvar el empleo fructífero de la lógica clásica en contextos en los que hacerlo resulta contraintuitivo. Para ello se han movilizado nuevos recursos matemáticos y conjuntistas en favor de la semántica abstracta, y los logros, en particular por lo que hace relación al desarrollo de teorías del punto fijo de la verdad y las definiciones inductivas, han sido ciertamente notables. Tales estrategias supervaluacionales han sido y son, de hecho, comunes a tratamientos de ambas paradojas en contextos «clásicos». [McGee 91] y

[Tappenden 93] han sistematizado explícitamente el método supervaluacional como estrategia común para abordar ambas paradojas. La idea básica es la siguiente: tanto el predicado de verdad como los predicados vagos se conciben como predicados parcialmente definidos, esto es, dotados de extensiones y antiextensiones «deficientes» o no exhaustivas, que son discrecionalmente extendibles por sus usuarios, bajo determinadas constricciones impuestas por el significado de tales predicados. Secuencias transfinitas de tales extensiones aseguran puntos fijos para el predicado parcial de verdad, mientras que secuencias de precisiones aseguran valores bivalentes de verdad para cada enunciado con predicados vagos. Por supuesto, distintas teorías supervalucionales definen tales secuencias de distinto modo.

Los tratamientos supervalucionales han de afrontar dificultades muy serias, dos de las cuales importa destacar aquí someramente con objeto de motivar la alternativa propuesta en el trabajo. Basta asumir que *parte* de la extensión de algunos de los predicados en cuestión sí está bien definida (funcional y bivalentemente) para que el lenguaje en el que el tratamiento supervaluacional se formule quede infectado con paradojas «de segundo orden». Si *P* es un predicado vago, el predicado "... es claramente una instancia de *P*" puede, a su vez, ser vago. La vaguedad no es vagancia para completar un proceso de precisificación, por cuanto en muchos casos tal proceso no puede siquiera comenzar (el punto se argumenta con detalle en [Salto 92] y, con más éxito, en [Williamson 94]). El caso del predicado de verdad está atestigüado desde [Kripke 75] y está profusamente discutido en la literatura. El segundo problema atañe a la imposibilidad práctica de aprender y emplear lenguajes que, libres de paradojas, requieren cantidades infinitas de extensiones y/o precisificaciones de los predicados generadores de paradojas. La estrategia supervaluacional requeriría ser complementada con una teoría clásicamente consistente capaz de justificar la realizabilidad de supertareas *semánticas*, y de justificar que de hecho las realizamos.

Este trabajo pretende fundamentar fehacientemente algunas y razonar someramente otras de las siguientes hipótesis:

- (a) La extensión de cada predicado sorites es una colección finita sobre la que no es posible definir un buen orden. Por tanto, su cardinalidad es indeterminada.
- (b) La hipótesis (a) aporta una solución natural a paradojas sorites y es compatible con, entre otras, la lógica clásica.
- (c) Colecciones finitas cardinalmente indeterminadas son capaces de simular por medios estrictamente finitarios los objetos matemáticos infinitos requeridos usualmente en el estudio de la semántica abstracta de las paradojas de la verdad

1

Las paradojas sorites nos retan a comprender cómo secuencias *finitas* de objetos o sucesos pueden no tener *término*. La secuencia de segundos que duró la infancia de la persona que lee esta página es obviamente finita, si ya no es un niño. Sin embargo, es imposible determinar qué segundo fue el último segundo de la infancia de quien esto lea.

Las opciones que aquí merece la pena considerar son las siguientes: ¿es inteligible y fructífero pensar en colecciones finitas que no tienen un último miembro?, o bien ¿es inteligible y fructífero pensar en colecciones finitas cuyo último miembro no puede ser determinado?

Que podemos comprender secuencias finitas de *números* que no tienen término en el sentido de ser indeterminadamente grandes, es algo que resultados metamatemáticos nos aseguran [Mycielski 86,92; Cardone 97]. En particular, conjuntos finitos indeterminadamente grandes se comportan, respecto a cualquier valor numérico dado efectivamente en tiempo real, exactamente como conjuntos infinitos. Ello permite, de hecho, *simular* con medios exclusivamente finitos axiomas análogos al de infinitud y, en suma, el propio continuo real.

La metáfora del zilión [Lavine 94] ilustra con fortuna la idea básica. Como el millón, el billón, ..., el zilión es un ordinal finito que comparte con otros n -llones la propiedad de ser mayor que cualquier número natural expresado por un numeral efectiva o explícitamente dado o inscrito en, digamos, cualquier sistema de notación instanciado en el planeta Tierra hasta el segundo milenio de nuestra era. Lo que distingue al zilión es el hecho de ser a la vez finito y «cardinalmente vago» o indeterminadamente mayor que cualquier número natural particular explícito. Parte de algunos argumentos ulteriores de este trabajo dependen de la inspección del modo como esta simulación toma cuerpo formalmente. Por este motivo se indica y comenta su base axiomática.

En rigor, para obtener una simulación finitaria $\text{Fin}(T)$ de cualquier teoría T de primer orden, procedamos¹ a incluir en su lenguaje una cantidad finita arbitrariamente grande de nuevos predicados $\Omega_0, \Omega_1, \dots$ (los subíndices son racionales, esto es, pares de números naturales y por tanto objetos finitos), y añadir como axioma cada instancia de los esquemas:

$$A1 \quad \Omega_p(a)$$

para cada término constante a del lenguaje recién enriquecido en el que no aparezcan predicados Ω_q , si $q > p$. Nótese que $\Omega_p(a)$ es un teorema de $\text{Fin}(T)$ para cualquier constante a del lenguaje de T y cada subíndice p . El sentido de $A1$, a la luz de las consideraciones anteriores, es naturalmente que cualquier numeral del vocabulario de T expresa un número en la extensión de Ω_p . Por

¹La axiomatización es una simplificación de la simplificación que {Lavine 94} hace de una versión de los axiomas originales de Mycielski.

tanto, asumiendo en lo que sigue que ω_p expresa la cantidad de elementos en la extensión de Ω_p , $a < \omega_p$, para cualquier numeral a del lenguaje de T . Indeterminadamente grande significa, en la presencia de $A1$, más grande que cualquier numeral inscrito efectivamente en alguna fórmula de T . Análogamente, $A1$ implica $\omega_p < \omega_q$ cuandoquiera que $p < q$. El esquema siguiente generaliza este hecho:

A2 Si $p < q$, entonces $\forall x_1 \dots x_n (W_p(x_1 \dots x_n) \rightarrow W_q(f(x_1 \dots x_n)))$

para cada símbolo n -ádico de función del lenguaje de T . Finalmente, el esquema siguiente asegura que sucesivos predicados no introducen explícitamente nuevos números mayores que ω_0 haciendo, indeseablemente, que ω_0 no sea indeterminadamente grande. En consecuencia, desde el punto de vista de cualquier número explícito en un numeral del lenguaje, todos los predicados nuevos son indiscernibles:

A3 $\forall x < \omega_p (\forall x < \omega_q A \leftrightarrow \forall x < \omega_r A)$

para $p < q$, $p < r$, $q, r < s$ para cada Ω_s que aparezca en A , que es una fórmula todos cuyos cuantificadores están ligados por predicados Ω .

El siguiente teorema [Mycielski 86, 92] es él mismo demostrable finitariamente y da contenido preciso al discurso acerca de la simulación finitaria de conjuntos infinitos:

$T \vdash A$ syss $\text{Fin}(T) \vdash A'$

donde A' es el resultado de acotar los cuantificadores en A con los nuevos predicados finitarios Ω , y éstos sólo aparecen en A' como cotas de cuantificadores.

$\text{Fin}(T)$ es por tanto una copia sintáctica de T , y ambas son a fortiori equiconsistentes. De hecho, T es consistente syss todo subconjunto finito de $\text{Fin}(T)$ es finitamente satisfacible. Por supuesto, las diferencias semánticas son notables, ya que $\text{Fin}(T)$ tiene modelos en los que los predicados Ω tienen dominios finitos indeterminadamente grandes syss T es consistente.

2

¿Significa esto que ω_0 es la respuesta a la pregunta por el número de segundos de la infancia de la persona que lee esto? No, en absoluto.

Retóricamente, ω_{77} sería una respuesta igualmente fundada, sin que nada en los axiomas que definen Ω_0 y Ω_{77} permita asegurar que no son *la misma* respuesta. Pero si no son la misma respuesta, esto es, si no expresan una y la misma cardinalidad, entonces la cantidad de segundos de la infancia no es conjunto, sino algún otro tipo de colección. Los axiomas de Mycielski imponen constricciones sobre el significado de “indeterminadamente grande” que aseguran la equiconsistencia de esta noción con la teoría estándar infinitaria de conjuntos. Pero manifiestamente, desde un punto de vista lógico, se limitan a indicar condiciones necesarias y suficientes para eliminar cuantificadores de T . En consecuencia, dejan abierta la posibilidad de

interpretar epistémica u ontológicamente la indeterminación de cada Ω . De ello se deduce que sólo la ulterior interpretación de los predicados Ω puede asegurarnos que son literalmente conjuntos capaces de expresar números naturales estándar. (Por otra parte, el problema metalógico de la consistencia - no relativa- de $\text{Fin}(T)$ también depende de ulteriores interpretaciones de tales predicados, como se señala en [Potter 96]).

En suma, los predicados vagos en general y los Ω -predicados en particular admiten ser entendidos: como predicados cuyo dominio finito es un conjunto con un último elemento inaccesible, o como predicados cuyo dominio finito es una colección finita que, literalmente, no tiene un último elemento (no está bien ordenada). Claramente intuiciones filosóficas radicalmente distintas motivan una u otra interpretación² (nota), y el argumento siguiente muestra una importante deficiencia en la interpretación (a).

Consideremos un conjunto finito arbitrario X de todos los números accesibles³ epistémicamente, con independencia de si satisface o no los axiomas de Mycielski. La interpretación (a) está comprometida con la existencia de tales conjuntos, por cuanto como tales interpreta los predicados Ω . El problema es el siguiente: al ser un conjunto finito, tiene un último elemento accesible l . Si la función s de sucesor es una función total sobre los números naturales accesibles, entonces, al ser X un conjunto estándar de números naturales, $s(l)$ es un número accesible. Lo cual contradice la maximalidad de l . Pero, por otra parte, si la función s no es total en X , entonces no tenemos seguridad alguna de alcanzar l con finitas aplicaciones de s . Ello contradice, por supuesto, la hipótesis de que X es un conjunto estándar de números naturales. Obsérvese que esta dificultad es independiente de que X satisfaga o no los axiomas de Mycielski. El problema no es que no pueda reconocerse que el conjunto de los números accesibles no es accesible. El problema es compatibilizar ese hecho con que tal conjunto sea finito en el sentido conjuntista de tener un último miembro.

Adicionalmente, quien, como [Lavine 94], asuma la interpretación (a) de los predicados Ω , ha de asumir que todos sus fragmentos también son conjuntos finitos estándar. Pero entonces de nuevo l es el último número accesible. Esto es, alguna secuencia finita de aplicaciones de s lo accede. Puede hacerlo de dos maneras. O bien demostrando que no hay numeral accesible mayor que l , o bien demostrando que todos los demás numerales

² Hay finitismos fútiles y finitismos eliminativistas. La sola existencia de contrapartidas finitarias de teorías arbitrarias plantea un metaproblema adicional al segundo problema de Hilbert: pueden, o no, considerarse superfluas las pruebas ideales que involucran objetos matemáticos infinitos, una vez que puedan, eventualmente, ser justificadas por medios finitarios.

³ Más exactamente, ω_0 accesibles, teniendo en cuenta que la secuencia misma de predicados Ω no es transitiva.

accesibles son menores que 1. La primera alternativa está vedada, al no podemos acceder a numerales mayores que 1. La segunda alternativa requiere recorrer todos los numerales menores que 1. Pero es imposible determinar cuántos son si precisamente no conocemos 1.

A continuación, ofrecemos un marco conceptual general en el que tiene cabida la interpretación (b). Un marco semejante se hace necesario por cuanto postulas colecciones infinitas que carecen de un último elemento resulta, a primera vista, implausible.

3

Consideremos por un momento la paradoja de Galileo. El hecho «paradójico» que haya tantos números naturales pares como números naturales se convierte, trivialmente, en pseudoparadójico sólo cuando se satisfacen las condiciones siguientes: (i) el mencionado hecho se plantea en términos conjuntistas, (ii) hay definiciones de teoría de conjuntos que cualifican debidamente la noción de ‘cantidad de elementos de un conjunto’. La paradoja se elimina reescribiéndola: por estipulación. Consideremos por un momento la colección $\{1,1\}$. ¿Qué cantidad de elementos tiene? Como conjunto, un solo elemento. Como multiconjunto, sin embargo, dos. Estas simples consideraciones muestran algo importante. Ilustran cómo para determinar numéricamente la cantidad de elementos de una colección arbitraria es necesario considerar de qué tipo de colección se trata. Es decir, para que conjuntos o números expresen cantidades, la noción «bruta» de cantidad ha de ser debidamente cualificada..

Por “cualificar debidamente” se entiende cualquier expediente capaz de establecer adecuadas condiciones de unicidad y equinumerosidad, esto es, determinar qué cuenta como unidad y qué cantidades se consideran equinumerosas. Por ejemplo, habitualmente cualificamos la noción de ‘cantidad de elementos de un conjunto’ estipulando (a) cada elemento distinto es único y debe contar como un objeto, y (b) la biyectividad es condición necesaria y suficiente para considerar equinumerosas cantidades.

En caso contrario, las consideraciones con las que abríamos este párrafo no tendrían sentido. No es una cuestión de prurito lógico aquello que exige hacer explícitos supuestos como (a) y (b). Tienen, por el contrario, un alto valor informacional.

Cantidades en bruto no son magnitudes y, en consecuencia, no cualquier cantidad es numéricamente representable en el sentido que hemos introducido. Aunque ciertamente sí lo sean cantidades debidamente cualificadas. Llamemos *numerosa* a cualquier cantidad cualificada por algún criterio adecuado de unicidad y equinumerosidad (criterios que no tienen que coincidir necesariamente con (a) y (b)).

Sin embargo, una cosa es llevar a cabo una operación de cuenta y otra cosa muy diferente es llegar a un resultado expresable mediante un número

cardinal. Cualificar como numerosa a una cantidad no basta para asegurar que ésta sea representable numéricamente mediante algún número cardinal. Tomemos, por ejemplo, el universo conjuntista ZF (sin axioma de elección) en el que se ha cualificado debidamente la noción de "cantidad de elementos de un conjunto". Esta cualificación no basta para asegurar que la numerosidad de cualquier conjunto ZF sea representable mediante un número cardinal. En consecuencia, no toda cantidad es numerosa, ni toda cantidad numerosa es representable mediante un número cardinal. Diremos que una cantidad es *cardinalmente numerosa* syss hay un único número cardinal capaz de representar su numerosidad. Por ejemplo, la cantidad de ecuadores terrestres es cardinalmente numerosa. Es claro que no toda cantidad es cardinalmente numerosa, por cuanto la cantidad de conjuntos, la cantidad de espacios vectoriales, la cantidad de ordinales, y otras muchas cantidades no lo son. Decimos análogamente de una colección arbitraria que es cardinalmente numerosa syss la cantidad de sus componentes es cardinalmente numerosa. Todo conjunto que pueda ser bien ordenado es cardinalmente numeroso. Pero no cualquier colección es cardinalmente numerosa, por cuanto no cualquier colección es un conjunto

4

Un *conjunto* es finito syss su cardinalidad es menor que $\aleph_0$. Ahora bien, si una colección no es un conjunto, entonces no puede ser finita en ese sentido. En particular, cualquier colección que no sea cardinalmente numerosa o no sea numerosa puede ser, en ese sentido, finita. Estas afirmaciones deben ser obvias, por cuanto la definición conjuntista de "finito" no es una definición simpliciter, sino más bien una definición de "conjunto finito". Resultará conveniente llamar **c-finita** a una colección precisamente cuando sea un conjunto finito. Por lo mismo, llamaremos **i-finitas** a todas y sólo las colecciones finitas que no sean cardinalmente numerosas (nótese que las colecciones finitas no numerosas son i-finitas).

Es tan difícil dar un significado preciso y no controvertido a la cuestión "¿hay colecciones i-finitas?" como a la cuestión "¿hay conjuntos infinitos?". En estas condiciones, más difícil aún es ofrecer una respuesta incontrovertida. Sin embargo, podemos exigir para ambas cuestiones la aplicación de los mismos estándares de rigor. Estos estándares permiten presumir la consistencia, y aún, en general, la verdad, del axioma de infinitud. Ahora bien, puesto que $\text{Fin}(T)$ tiene modelos no conjuntistas (de acuerdo a las definiciones arriba ofrecidas) del axioma de i-finitud, el teorema de equiconsistencia de Mycielski asegura la consistencia relativa del axioma de i-finitud respecto del axioma de infinitud⁴. Con independencia de ello, sin duda sólo razones adicionales pueden hacer plausible el axioma de i-finitud. Pero

⁴ Cfr. [Vopenka 79] y [Salto y Méndez, 99].

estas razones existen: las propias paradojas sorites avalan intuitiva y naturalmente el axioma de i-finitud. Las secuencias sorites son finitas y no tienen un último término. Reconocer que tales colecciones no son conjuntos es tan fructífero como reconocer que la colección infinita de los números naturales sí lo es.

5

La plausibilidad con la que, a primera vista, cabe afirmar:

j. La cantidad de segundos de la infancia del lector es finita

jj. No hay un último segundo de la infancia del lector

bien pudiera ser refutada. ¿Qué podría refutarla? No parece que el descubrimiento de algún hecho acerca del envejecimiento humano pudiera refutarla. Presumiblemente existen intereses lingüísticos que avalan y explican el significado convencional de los términos involucrados, de manera que tampoco parece que tales convenciones se alteren. Es dudoso incluso que, aun si se alterasen, ello fuese relevante a las ocurrencias de enunciados arriba expuestas. De hecho, las razones por las que la comunidad filosófica rechaza en general esta aproximación directa a la paradoja sorites son de carácter puramente lógico.

Las razones lógicas aducidas están basadas en inferir erróneamente de jj secuencias de enunciados en la forma:

k es un segundo de la infancia del lector $\supset$ k+1 es un segundo de la infancia del lector

De tales secuencias indeseables o de conjuntos de ellas se deducen sin duda falsedades patentes. Bien por inducción, bien por finitas aplicaciones de Modus Ponens. No se ha apreciado que el condicional material en cuestión no se deduce de jj, como tampoco que las colecciones i-finitas no pueden estar cerradas inductivamente. Finalmente, tampoco se ha apreciado que sólo una cantidad i-finita de aplicaciones de Modus Ponens permitiría recorrer una secuencia i-finita. Las pruebas, sin embargo, son precisamente objetos matemáticos paradigmáticamente finitos.

BIBLIOGRAFÍA

- Carbone, F. "Strict Finitism and Feasibility" en: Leivant, D. (ed) *Logical and computational Complexity. Selected Papers*. Springer, 1995, 1-21
- Lavine, S. *Understanding the infinite*, HUP, 1994
- Kripke, S. "Outline of a Theory of Truth", *Journal of Philosophy*, 72 (1975), 53-81
- McGee, V. *Truth, Vagueness and Paradox*, Hackett, 1991
- Mycielski, J. "Locally finite Theories", *Journal of Symbolic Logic* 46 (1986), 59-62
- Mycielski, J. "The Meaning of pure Mathematics", *Journal of Philosophical Logic* 18 (1989), 315-20
- Mycielski, J. "Quantifier-free versions of first order logic and their psychological significance", *Journal of Philosophical Logic* 21 (1992), 125-47
- Peña, L. "Lógicas Multivalentes" EIAF, 1996.
- Potter, M. "Taming the Infinite" *British Journal of Philosophy of Science* 47 (1996), 609-19

Salto, F. *Identidad Numérica*, tesis doctoral, Universidad de Salamanca, 1992

Salto, F. y Méndez, J.M. "Set-free completeness" enviado, 1999.

Tappenden, J. "The Liar and Sorites Paradoxes: Toward a unified Treatment", *Journal of Philosophy* 90 (1993), 551-577

Vopenka, P. *Mathematics in Alternative Set Theory*, Teubner Verlag, 1979

Williamson, T. *Vagueness*, Routledge, 1994

Francisco Salto Alemany
Universidad de Salamanca
e-mail: Salto@gugu.usal.es





ON LOGICAL FORM

ELENA TATIEVSKAIA
Augsburg University

ABSTRACT: Russell's notion of proposition gives rise to the idea of logical form as the structure of proposition. Identification in the notion of proposition of the object of cognition with the content of knowledge implies the idea of objectivity of logical forms. Further analysis of the notion of proposition makes this idea questionable. The change in the treatment of proposition (the theory of descriptions and the theory of multiple relations) makes the logical form a form of propositional cognition.

KEYWORDS: Complex, logical form, proposition, fact, propositional attitude.

1

Russell's interest in foundations of mathematics as well as his attempt to answer anew Kant's question of the nature of mathematical judgements leads to formulation of the notion of proposition. This notion is necessary to differentiate the subject and the content of pure mathematics from the mental device by means of which that subject is comprehended and that content is formulated - from a judgement being a subjective mode of receiving truth. Proposition is required as "a name for the true or false as such"¹. The analysis of this notion calls forth the problem of the structure of proposition and of the nature of its constituents.

Proposition is for Russell an objective structured complex. The kinds of this structure characterize types of propositions and thus enable the definition of mathematics and determination of its nature. Less immediate consequence of differentiation of the kinds of propositional structure is the categorial classification of things that can occur in propositions as their constituents and further ontological description of the world on the basis of such a classification.

¹ Russell, Bertrand: *The Principles of Mathematics*, London: George Allen & Unwin, 1956, p.XIX.

Besides constituents - terms or objects (everything that can be a subject or of which something can be predicated and what can be counted), propositions as structural units possess a form - a mode of combination of these constituents. This form is a constant revealed in a complex due to its analysis.

Proposition can be analysed in two ways - by means of propositional function and by means of the notion of relation. In spite of advantages of the formal applicability of the concept of propositional function, the analysis by means of it cannot specify the analyzed complex really, since it cannot determine the position of each constituent in the complex, i.e. the sense of relation connecting the constituents of propositions. For instance, the formalized expression of the proposition *The Mont Blanc is higher than any other European mountain* based upon the usage of propositional function is indiscernible from the proposition *Any European mountain is higher than the Mont Blanc*. The propositions are obviously distinct, but the corresponding propositional function is one and the same - x is higher than y or, if we substitute a variable for *is higher than*, xRy (or $R(x,y)$). Such a substitution, in order to yield meaningful propositions as values of a propositional function should be restricted. The range of the variable R should be different from the ranges of other constituent variables. Else it will be impossible to pick out something constant in the proposition that can be the basis of the ranking of this one together with other propositions of the same type, whereas such a ranking should be the result of analysis of the propositions by means of propositional function.

Another way of analysis by means of the notion of relation taking into account not only the mode, but also the order of combination of constituents, is meaningful, first of all, from the philosophical point of view and has nearly no use in the field of realization of the program of logicism.

Whichever of these two methods of the analysis of proposition is taken as a fundamental one, it is possible to speak about the form of a complex as about its logical form. First of all, of whichever type the given proposition is, whether it consists in the relation of a term to a class a member of which the term is, or in a relation between two or more terms, or in a relation between two or more classes (of terms), or in a relation between propositions themselves, - all these notions are notions defining logical constants². Secondly, various types of the structure of propositions have general and constant character. Such a character implies universal validity of logical principles. Thirdly, the structure of propositions belongs to characteristics of their general nature which is studied as a subject of logic in contrast to psychology and the theory of knowledge³.

² Op. cit. note 1, § 1

³ Russell, Bertrand: "Meinong's Theory of Complexes and Assumptions". *Mind*, vol. 13, N 49 - 52, 1904, p.353.

One of the above facts, namely the fact that the logical forms possess universal validity, may be caused either by the objectivity of logical forms and principles themselves, or by the correspondence between them and some forms or laws of objective world.

Before "On Denoting" (1905) Russell identifies the object of cognition with the content of knowledge. This identification is necessary to guarantee the objectivity of knowledge which for Russell is indeed equivalent to intersubjectivity. Proposition as a complex is a unity of a form (a structure) and a content (its constituents). Propositions according to Russell are objective. Their being does not depend upon their being cognized by somebody. They are not created by the cognizing subject, but are what he conceives. The objective nature of propositions means the objectivity both of their form and their content. But this objectivity is questionable.

The logical form defined as the structure of proposition has no special linguistic expression in the sentence which means a proposition. That is, there is no special linguistic sign of the ordinary language which stands for the logical form. Clearly, that every sign should have some objective meaning and thus should stand for something in the world⁴, does not imply that everything objective must be represented by some sign. Such a sign can be arbitrarily invented in order to speak about objects which are not daily discussed and which therefore are designated otherwise than things, concepts and propositions. But the lack of such a sign nevertheless may indicate the subjective character of the logical form.

The logical form or the structure of proposition is essential as far as it determines the content (the intension) of a sentence - the proposition. Already the simple change of order of constituents of a sentence gives a completely new proposition whose truth-value can be opposite to the truth-value of the initial one as the above example with the Mont Blanc shows. What is the reason of this linguistic change? If this reason lies in the objective state of affairs, then propositions are given to the cognizing subject as building up two equally independent kinds - true and false propositions. Their truth and falsehood is in such a case their original property supposing no dependence upon other entities. If, on the other hand, the above reason lies in the subjective activity, then propositions should have subjective character. If such a subjective character is not accepted (what takes place in Russell's works at least before 1905), there exists another possibility - to search for the kind of objectivity peculiar to logical forms and being other than objectivity ascribed to propositions and their constituents.

⁴ Op. cit. note 1, § 99.

2

Russell classifies logical forms and principles as objects which have to be cognized. One consequence of such a treatment is the mentioned above consideration of truth and falsehood of propositions as properties of objects of cognition, not of cognitive activity or its results.

In 1902 Russell discovers the antinomy of propositions⁵, connected with predicating any kind of properties, among them of truth or falsehood, of propositions, further classification of such propositions and predicating the property of the class. Such a predication takes place again in the form of proposition. Any attempt to determine, whether this proposition itself has the property which it predicates of the formed class of propositions and therefore belongs to this class, is either meaningless or leads to a contradiction. The ways to escape contradictions of such a sort are, firstly, to forbid judgements concerning the totality of propositions, and secondly, not to treat propositions as independent objects whose properties have their origin alone in the nature of propositions themselves. There exist following reasons in favour of the second alternative or of its combination with the first one.

If propositions are initially true or false, and both of them are given to the cognizing subject and can be perceived or conceived by him, it is not clear how he can estimate their truth-value. What are the criteria of truth of cognized propositions? It cannot be only the evidence of the cognized, since both kinds of propositions should be equally evident. In order not to accept the correspondence theory of truth, which makes doubtful the objective nature of propositions, the idea of the consistency of the cognized with the present content of knowledge can be suggested. But the same question concerns this very already existent content which can happen to be incorrect as it often takes place in the history of the development of scientific ideas.

Not only the idea that truth and falsehood are properties of objects of cognition, but also Russell's semantic conceptions concerning propositions are problematic. First, problematic is the relation between a proposition which involves an exemplification of a concept (for example, *I met a man*) and the sentence standing for such a proposition. Such a sentence occurring in the process of intercourse between two (or more) persons designates indeed two propositions. They differ from each other as to their structure and constituents, and this difference depends upon the role of the subject in the intercourse - whether he is the author of the sentence acquainted with one of the propositions immediately, or the recipient who is informed about it. For instance, the proposition mentioned above *I met a man* contains from the point of view of speaker the author of the corresponding sentence, his relation to some particular person and this person itself. For his interlocutor the proposition involves the author, his relation to some variable that belongs to

⁵ Op. cit. note 1, § 500.

the extension of some concept (here it is *man*) and an additional proposition including the relation between the variable and the concerned concept or rather its extension. Thus one and the same sentence designates a relation between two particular objects and simultaneously a relation of a particular object and an indefinite part of the extension of a concept which in its turn itself can be treated as a concept. The question is: which of these two complex objects (propositions) is really objective? If one of them is a genuine object in comparison with the other, then a definite part of propositions (perhaps the greatest part if estimated as a real object is the complex with which the author is acquainted) is deduced in some way from the others and therefore cannot be treated as involving objective entities.

Another problem is connected with the structure of false propositions. If a true proposition has a relational structure and consists in some relation between, for instance, two terms (let such a proposition be *The Mont Blanc is higher than the Zugspitze*), it is natural to suppose that a false proposition having the same structure, such as *The Mont Blanc is higher than the Everest*, does not indeed involve the relation *is higher than*. But what are in that case the constituents of the false proposition? If there is no relation *is higher than* between Mont Blanc and Everest taken in the above order, then there is no complex object consisting in the relation between these two objects. We can only suppose that the false proposition has some other status than the true one.

The case of false propositions and coinciding with the common sense idea that sentences designating them possess them as their meanings, but have no denotation, lead to the supposition that plausible is assumption of another kind of objective reality than that of propositions.

The notion of proposition is revised. This revision is realised in the theory of descriptions (1905) and the theory of multiple relations (1907 - 1913).

The first of them treats propositions as "incomplete symbols". They are assemblages of terms discovered during the analysis of perceived objects - facts. Judging the subject synthesizes these terms and in such a way they are transformed into the whole - the meaning of a sentence, or proposition, which is still considered to be different from its linguistic form (a sentence), but whose truth-value (and therefore the truth-value of the appropriate sentence) depends upon existence of the fact which is perceived and analyzed.

The perceived fact is a unit whose structure is discovered and asserted by the subject. As far as the fact in perception is yet undivided and unstructured, the perception itself is a binary cognitive relation of the subject to the fact. Judgement as another kind of cognition is a multiple relation between the subject and the constituents of proposition, to wit constituents picked out by the subject in the process of analysis of the presentation of a fact. Thus, each unit of cognition different from perception is interpreted as propositional attitude (an attitude of the subject to a proposition as the

assemblage of possible constituents of the fact). The theory of multiple relations tries to define the form of such an attitude.

3

Taken in the broad sense, propositional attitudes of different kinds are diverse cognitive complexes (or facts) with exception of acquaintance (or perception). Peculiar to acquaintance is the simplicity of its object. Proposition can hardly be called "the object" of propositional attitude. Propositions differ from the objects of acquaintance in their being true or false even if the propositional attitude of a subject to a proposition does not include the assertion or negation of the last. That propositions are not admitted of being objects saves from the assumption of objective reality of false propositions. The false propositions as well as the true are non-entities and exist only as meanings of sentences formulated as the result of some propositional attitude.

Each propositional attitude as a complex has a definite property - its form⁶. The position that each constituent of the complex takes in this form is the property of an appropriate constituent.

Each cognitive complex involves the cognizing subject, the constituents of a proposition, the form of the proposition and a propositional attitude in a narrower sense of the word, namely, a definite kind of the subjective attitude to the proposition - either understanding, or belief, or disbelief, or doubt, or any other possible sort of cognitive relation. Here it must be noted that besides the form of proposition in the case of the relational proposition one more constituent occurs in the cognitive complex. This constituent is the difference between the logically possible complexes and the real one. The point here is the notion of the sense of relation formulated in Russell's correspondence with Frege, *The Principles of Mathematics* and some further works. For instance, if we take the proposition *The Mont Blanc is higher than the Zugspitze*, its form may be symbolized as "xRy" if for the name of every constituent some variable is substituted. This symbolical expression corresponds really to two logically possible propositions - the original one and *The Zugspitze is higher than the Mont Blanc*. Only one of these logically possible complexes is real and usually only one of them is believed to be real and therefore only one is further asserted or negated. All these constituents of the cognitive complex possess some position in it. Therefore it is possible to speak about two kinds of form. Firstly, about the form of a proposition, and secondly, about the form of a cognitive complex.

⁶ Russell, Bertrand: "Theory of Knowledge: The 1913 Manuscript". Eames, E.R., Blackwell, K. (eds): *The Collected Papers of Bertrand Russell*, vol. 7, London, Boston, Sydney: George Allen & Unwin, 1983, p.81.

Both of these kinds are defined as logical forms. Logical forms of propositions are considered to be already accepted as such. Logical forms of cognitive acts or of conception of propositions demand such an acceptance and should enlarge the existing inventory of logical forms.

The idea that logical form is peculiar to a propositional attitude (to a judgement or to the act of understanding) calls forth the question: what is the form in its both exemplifications? Is it some object or a kind of conception of complexes existing in the world?

Considering the problem of logical form in 1912 as an object of possible theoretical studies, Russell suggests two ways of its treatment and definition. The logical form may be, on the one hand, assumed as something being but not definable and therefore taken as a primitive notion. It can be described as opposed to constituents of a complex as non-constituent. It is "the way the constituents are put together"⁷.

On the other hand, defined may be the relation of *having the same form* that exists between two complexes. This relation holds when one complex "can be obtained from the other by mere substitution of new terms in other places". But a simple attention shows, that substitution takes place only as regards the symbols for complexes or their form. If the complex name standing for a complex is considered as the result of analysis⁸, a symbol containing variables can be treated as a result of the formal analysis of the complex, or the analysis by means of which the form of the complex (the mode of combination of its constituents) is determined⁹. Thus, when we determine the relation of having the same logical form between any two complexes, we compare complexes by means of comparison of symbols standing for them and therefore indeed we compare the results of analysis carried out by us in the process of cognition. But such an interpretation makes the logical form the property of subjective cognitive activity and entails the assumption that the logical form itself is rather a symbol than something objective, whether this objective is thing, or property, or relation, or some object of different kind. This assumption in its turn leads to the question, what it is, for what such a symbol stands and the whole problem regarding logical form or its objective prototype arises afresh. Another problematic peculiarity of the relation of having the same form consists in the fact that when the analyzed complex includes some asymmetrical relation, *having the same form* is defined on the basis of the notion of form, since the position of each term in the complex should be mentioned.

⁷ Russell, Bertrand: "What is Logic?". Slater, J.G., Frohmann, B. (eds): *The Collected Papers of Bertrand Russell*, vol. 6, London and New York: Routledge, 1992, p.55.

⁸ Op. cit. note 6, p.128.

⁹ Op. cit. note 6, p.119.

All these problematic conclusions are what Russell would like to escape. He is apt to treat logical forms rather as objects than as symbols or properties of subjective analysis. One of his reasons for such a treatment lies in the mode of existence of logical forms. On the one hand, logical forms exist as the mode of combination of constituents of complexes. As such they can be objects of cognition, in particular of understanding, and being involved in a complex are subject to analysis. On the other hand, logical forms exist as independent objects to which definite propositions or sentences correspond, for instance such as "Something has some predicate" or "Something is somehow related to something". Therefore logical forms themselves can be objects of cognition. But they cannot be analyzed - they involve no constituents, but only variables. Logical forms seem to have a structure, but they do not have it, since logical form is rather itself a structure¹⁰. Two consequences of this fact: first, there must be a special kind of cognition perhaps including no analytical activity, second, there arises a question, whether some propositions really correspond to such objects as logical forms. The first problem is discussed by Russell. Logical forms are declared to have a specific form of objectivity. They are facts having a special nature, namely being simple¹¹. This nature of logical forms requires a dual cognitive relation which resembles the acquaintance in which the object is not analyzed and which is devoid of properties of being true or false. This cognitive relation is often called understanding which in the case of logical forms is not differentiated from the acquaintance as described above. The acquaintance with logical forms is defined as "logical intuition"¹². It is the characteristic of the mode of their cognition.

The supposition that logical forms are facts seems to remove the problem of existence of propositions corresponding to them. But this problem is indeed essential. First, the proposition is a complex. If the logical form is simple and cannot be as such analyzed, there cannot exist any complex symbolizing it. Therefore the sentence standing for the logical form is not indeed a sentence but a complex name involving nothing but linguistic variables. Such a sentence as a genuine proper name has no meaning, but only denotation - the logical form. Logical principles which assert or negate definite relations between logical forms and determine universal or particular truth of complexes involving logical forms, are expressed in the sentences whose truth-value may be estimated and which allow to speak about propositions corresponding to them. In such a case we can treat logical forms as a kind of particular things which are universal in the sense that they define classes of propositions being of the same logical form and have no definite

¹⁰ Op. cit. note 6, p.114.

¹¹ Op. cit. note 6, p.129.

¹² Op. cit. note 6, p.101.

space- or time position, but nevertheless play the same logical role as particular things play in propositions other than logical. Namely, they are only subjects of predicates and terms of relations. The problem here is following. The sentence "If something (an x) has some relation (an R) to something (an y), then for all these x and y something (an y) has some relation (an Q) to something (an x)" asserts in the case of the acceptance of the above conception not the relation of implication between two complexes or propositions, but such a relation between two terms. As far as the relation of implication is traditionally defined as a relation between propositions, it is necessary either to define this relation anew, or to assume the existence of another relation having the same linguistic cover but being different from the relation of implication. Such a deep revision of logical theory could hardly be Russell's aim.

Thus, Russell's thesis according to which logical forms are objects given to the cognizing subject in experience seems to be doubtful. His theory rather ascribes another kind of reality to them.

Firstly, understanding of forms in spite of its closeness to acquaintance belongs to the step of cognition that follows not only the acquaintance, but also propositional cognition.

Secondly, an isolated understanding of a logical form is more difficult than understanding of complexes being examples of some of the forms. One of the reasons: pure logical forms "are fugitive to attention"¹³ which accompanies the perception. This idea is close to Meinong's concept "wahrnehmungsflüchtige Gegenstände" used to speak about objectives or objects of higher order (Gegenstände höherer Ordnung) that have the psychical nature¹⁴. This analogy can lead to conjecture that Russellian logical forms belong in fact to instruments of comprehension of facts.

Thirdly, the logical form of complexes is part of cognitive acts. Logical forms are used by subject when he judges. Particularly, judging the subject asserts some relation between constituents of the proposition and one of the possible forms that can unite these constituents.

Fourthly, the logical form is stated due to analysis of a series of complexes which is based upon formalization of signs representing the complexes and their comparison. Since every analysis and formalization includes some moment of arbitrariness, it is advisable to escape treatment of the logical form in all cases as a given object.

Fifthly, the assumption that logical forms are a special kind of facts in spite of its correctness in definite fields of knowledge, such as, for instance,

¹³ Op. cit. note 6, p.132.

¹⁴ Meinong, Alexius: "Über Gegenstände Höherer Ordnung und deren Verhältnis zur Inneren Wahrnehmung". Haller, R., Kindinger, R., Chisholm, R.M. (eds): *Alexius Meinong. Gesamtausgabe*, Band II, Graz – Austria: Akademische Druck- u. Verlagsanstalt, 1971, p.437.

logic and philosophy, can have undesirable consequences for the basic notions of these sciences.

All these facts make plausible the idea that the logical form is one of the forms of conception of the world and of facts constituting it.

REFERENCES

- Russell, Bertrand: *The Principles of Mathematics*, London: George Allen & Unwin, 1956.
- Russell, Bertrand: "Meinong's Theory of Complexes and Assumptions". *Mind*, vol. 13, N 49 – 52, 1904.
- Russell, Bertrand: "Theory of Knowledge: The 1913 Manuscript". Eames, E.R., Blackwell, K. (eds): *The Collected Papers of Bertrand Russell*, vol. 7, London, Boston, Sydney: George Allen & Unwin, 1983.
- Russell, Bertrand: "What is Logic?". Slater, J.G., Frohmann, B. (eds): *The Collected Papers of Bertrand Russell*, vol. 6, London and New York: Routledge, 1992.
- Meinong, Alexius: "Über Gegenstände Höherer Ordnung und deren Verhältnis zur Inneren Wahrnehmung". Haller, R., Kindinger, R., Chisholm, R.M. (eds): *Alexius Meinong. Gesamtausgabe*, Band II, Graz – Austria: Akademische Druck- u. Verlagsanstalt, 1971.

Elena Tatievskaja
Augsburg University

EL GATO ESTÁ SOBRE LA ALFOMBRA PERO YO NO LO CREO

JAVIER VILANOVA ARIAS
Universidad Complutense de Madrid

ABSTRACT: According to Austin, formal logic can't explain why "The cat is on the mat but I don't believe it" is self-contradictory, because "I affirm that p" implies "I believe p" in a non logical sense ("non logical" means that "I believe p" is still possible when p is affirmed). This paper tries to give a formal explanation of the sentence that shows the principles that make the sentence absurd. I will argue that these principles are not "semantic conventions of our language" (as Austin says), and that there are situations where the sentence doesn't drive to contradictions.

KEYWORDS: Modal logic, epistemic logic, knowledge, belief, assertion, J. L. Austin

Introducción¹

Este trabajo puede contemplarse en el horizonte de la tradicional polémica entre dos "ramas", escuelas o tendencias dentro de la Filosofía Analítica en torno a la utilidad y alcance de los instrumentos formales de análisis en el estudio de los lenguajes naturales y la resolución de problemas filosóficos. Frente a aquella corriente (Strawson, Austin, Ryle...) que considera a la lógica formal insuficiente para expresar todas las sutilezas del lenguaje natural, defenderemos la tradición iniciada por Leibniz y continuada por autores como Frege, Russell, Carnap o Hintikka, tradición que se caracteriza por la convicción de que el uso de herramientas formales puede ayudar a aclarar y resolver problemas filosóficos. Con esta polémica como trasfondo, acudiré a uno de esos pequeños rompecabezas con los que la tradición informal se solaza en detectar las sutilezas "informalizables" del lenguaje natural, y

¹ Este trabajo se encuadra dentro de una investigación más general cuyo objeto de estudio son lo que en Vilanova (1998) y Vilanova (1999) denominé "nociones epistémicas modales" (aquellos conceptos como "cognoscibilidad" o "inverificabilidad", cuyo ámbito temático es el de la epistemología o la teoría del conocimiento, y que poseen una indudable nota modal), y cuyo propósito último es proporcionar una teoría formal que de cuenta de las propiedades de tales nociones y sirva de instrumento de análisis de las distintas teorías y concepciones epistemológicas que las involucran.

procuraré mostrar en qué aspectos resulta iluminador el análisis formal: precisamente en la medida en que nos ayuda a averiguar los presupuestos y las consecuencias de nuestras tesis (o las de otros), así como a evaluar su corrección formal.

La tesis de Austin

En "El significado de una palabra" (Austin 1961) Austin presenta la siguiente "curiosa" oración:

Tomemos la bien conocida oración "El gato está sobre la alfombra y yo no lo creo". Eso parece absurdo. Por otro lado, "El gato está sobre la alfombra y yo lo creo" parece trivial. Si aceptásemos una dicotomía aceptada y dijésemos que o bien una proposición p implica otra proposición r , o p es perfectamente compatible con $\neg r$, al instante en nuestro presente caso nos sentiríamos tentados a decir que "El gato está sobre la alfombra" implica "Yo lo creo": de ahí no sólo la trivialidad de añadir "y yo lo creo" sino también de añadir "y yo no lo creo". Pero por supuesto "el gato está sobre la alfombra" no implica "Austin cree que el gato está sobre la alfombra" ni tampoco "el hablante cree que el gato está sobre la alfombra" -pues el hablante puede estar mintiendo. La doctrina que se produce en este caso es, no que p realmente sino que el aseverar que p implica "yo (quien asevera p) creo que p ". Y aquí a "implica" se le debe dar un sentido especial... (Austin p. 78)

Según Austin, la diferencia entre este sentido especial de implicación y el sentido de implicación lógica usual es que (p siendo aseverado), si p implica en el sentido lógico q , entonces "podría no q " es absurdo, mientras que si p implica en el sentido especial q , entonces "podría no q " no es absurdo:

La consideración de la oración "podría no q " suministrará una prueba tosca de si p "implica" r en el sentido ordinario, o en el sentido especial de "implica". (Ib., 78)

Desde luego, las oraciones del tipo " p pero no creo que p " habían llamado la atención de lógicos y filósofos mucho antes que Austin la utilizará en este artículo², pero es él quién más nítidamente la toma como elemento de juego dentro de la polémica entre promotores y detractores del uso de métodos formales en la investigación filosófica (también es Austin, dicho sea de paso, uno de los que más claramente señala la operación de aserción involucrada). Lo

² Versiones antiguas aparecen ya en las colecciones de *Sophismata* medievales. Sustituyendo la creencia por la sapiencia, Moore popularizó a principios de siglo una versión de esta paradoja, que sigue dando quebraderos de cabeza a los lógicos contemporáneos a través de la paradoja de Fitch. Véase Martínez, Sagüillo y Vilanova (1997), y Vilanova (1999) para una discusión del tema.

que Austin parece dar a entender es que el análisis lógico es incapaz de dar cuenta del sentido en que la afirmación de p implica la creencia de p , precisamente porque esta implicación depende no de principios sintácticos o deductivos, sino de convenciones semánticas *relativas al modo en que usamos las palabras* en situaciones (Ib., p.78). Hasta aquí, y acorde a mi forma de verlo, Austin lleva toda la razón excepto en un punto: cuando dice que el análisis formal es incapaz de dar cuenta de las “convenciones semánticas relativas al modo en que usamos las palabras”³. En mi opinión, Austin da por zanjado el asunto demasiado pronto, apelando a tales “convenciones”, con lo cual nos deja tan a oscuras como estábamos al principio. Pienso que si se procura (i) dar una traducción formal de su tesis, y (ii) comprobar a partir de qué principios (convenciones) resulta deducible, veremos la misteriosa oración sobre el misterioso gato bajo una luz más blanca.

Con respecto al primer punto (la traducción formal) me parece obvio que hemos de acudir a operadores epistémicos, o a predicados epistémicos aplicados a oraciones (no haré aquí mayores diferencias entre ambos). Resulta, al menos, necesario un operador de “creencia” que dé cuenta del hecho de que un agente epistémico guarda una determinada creencia en un momento concreto: “Cp” significará a partir de ahora “el sujeto X cree que p (en el momento t)”. Dado que Austin afirma que es no “ p ” a secas, sino “el aseverar que p ” lo que implica “Cp”, hemos de añadir también un operador de aseveración o afirmación: “Ap” significará “el sujeto epistémico afirma o se compromete con la verdad de p en el momento t ”. Teniendo en cuenta que el ejemplo involucra algunas tesis acerca de lo que el hablante sabe cuando afirma algo, o al menos lo que sabe sobre sus creencias, precisamos un tercer predicado epistémico: por “Sp” entenderé “el hablante o sujeto epistémico sabe o conoce que p en el momento t ”.

Una forma apresurada de plasmar el principio de Austin es tomar literalmente su afirmación de que el afirmar que p implica que se cree que p : $Ap \rightarrow Cp$. Ahora bien, si nos fijamos bien, lo que el ejemplo de Austin viene a mostrar no es meramente que se da el anterior principio, sino que sobre la base de ese principio **junto con convenciones semánticas de nuestro lenguaje,**

³ En mi opinión, si hay algo interesante que pueda hacer el análisis lógico por los lenguajes naturales, precisamente es dar cuenta de dichas convenciones. Los logros de la lógica temporal, la lógica modal, la lógica condicional, la teoría formal de las presuposiciones, etc... dan cuenta de las convenciones que rigen el uso de locuciones como “necesariamente”, “hasta ayer”, “si... entonces...”. Creo que no resulta exagerado decir que el propósito principal de quienes construyeron las lógicas proposicional y de predicados clásicos era precisamente dar cuenta de las convenciones que rigen el uso de palabras como “y”, “no”, “todo” o “algún”..

la afirmación simultánea de p y de la no creencia de p es absurda. Lo que defiende Austin es que, para cualquier fórmula p , suponer $A(p \wedge \neg Cp)$ lleva a una contradicción; o lo que es lo mismo, que para cualquier fórmula p , se da que $\neg A(p \wedge \neg Cp)$. Además, y este es un punto sobre el que Austin pone un especial énfasis, afirmar que p no implica la imposibilidad de creer que p : puede aseverarse que p si que ello excluya la posibilidad (pero sí la verdad actual) de no creerse p , de dudar de p . Esta es la segunda parte de la tesis de Austin: puede darse a la vez que afirmo p y que es posible que no creo que p . Reuniendo ambos *dicta* llegamos a la formulación de la:

TESIS DE AUSTIN (TA): para toda fórmula p ,

$$(i) \neg A(p \wedge \neg Cp)$$

y ello aunque para alguna fórmula q se dé:

$$(ii) Aq \ \& \ \Diamond \neg Cq.$$

Algunas nociones de lógica epistémica

A pesar de amplia nómina de detractores con que cuenta (algunos de ellos ilustres: Quine, Barwise, Perry...), ninguna otra teoría se ha mostrado tan eficaz a la hora de explicar nociones intensionales y dar cuenta de modalidades como la semántica de mundos posibles. En efecto, es bien conocida la inmensa capacidad expresiva de esta semántica, que va desde el nivel de abstracción de la semántica universal de Montague hasta el grado de concreción y complejidad de los desarrollos de la misma efectuados para la lógica condicional, la lógica monomótona o la lógica borrosa. Por otra parte, desde la semántica de mundos posibles se han efectuado satisfactorios acercamientos tanto a las modalidades, como a las actitudes epistémicas, lo que induce a confiar en el éxito de la combinación de ambos enfoques que el entrecruzamiento de nociones epistémicas y modales requiere. Además, una ventaja importante de la semántica de mundos posibles es su alta cogencia. En efecto, su plausibilidad intuitiva permite que además de averiguar las relaciones formales entre conceptos, también seamos capaces de “ver” y comprender por qué se dan esas relaciones, todo ello dentro de un marco conceptual muy próximo al constituido por nuestras propias percepciones intuitivas e introspecciones. En nuestro análisis explotaremos al máximo este carácter a la vez integrador e intuitivo de la semántica de mundos posibles.

En adelante hablaremos de un lenguaje objeto L formado por enunciados p, q , etc.. Hablaremos también de un conjunto de índices o mundos posibles $U = \{u, v, x, y, z, \dots\}$, y una función de interpretación $\#$: $L \rightarrow \text{Pot}(U)$ que asigna a cada enunciado p del lenguaje un subconjunto de U formado por todos y sólo aquellos mundos donde p es verdadera. En la ya clásica obra *Knowledge*

and Belief, Hintikka propone representar un corpus o estado de conocimiento mediante un conjunto de mundos posibles, a saber, el conjunto de mundos lógicamente compatibles con ese corpus de conocimiento. Denominemos al conjunto de mundos posibles compatibles con lo que un sujeto epistémico determinado sabe en un momento concreto t , la *situación epistémica actual*, y simbolizémosla como “ e^a ” (el término es mío). Aparte de la situación epistémica actual podemos tomar en cuenta otros posibles estados de conocimientos en t , aquellos que se hubieran producido si el agente hubiera llevado a cabo determinados procesos o acciones de adquisición de conocimientos que estaba en su mano hacer, y por los motivos que fueran no llevo a cabo. Sea $E = \{e_a, e_1, e_2, \dots\}$ dicho conjunto de situaciones epistémicas posibles en t (siendo e_a como antes la situación epistémica actual). Como se habrá adivinado, “ Sp ” es verdadero en el mundo actual si p es verdadero en todo mundo posible perteneciente a e_a . “ $\Diamond Sp$ ” es verdadero si para algún elemento e_i de E , p es verdadero en todo mundo posible perteneciente a e_i .⁴

Introduzcamos un concepto análogo al de situación epistémica para la creencia. Dado un instante de tiempo y un individuo determinados (en este apartado prescindiremos de los índices temporales), sea la *situación doxástica actual* d^a el conjunto de mundos posibles compatibles con las creencias del individuo en ese instante. Aparte de d^a , podemos tener en cuenta otros posibles estados doxásticos en el momento t , $D = \{d^1, d^i, d^k, \dots\}$. Representemos por “ C ” el predicado “... es creído por tal individuo en tal momento”. “ Cp ” es verdadero en el mundo actual si p es verdadero en todo mundo posible perteneciente a d^a .

Adviértase que de la definición anterior se sigue el principio de no

⁴ Resulta importante hacer notar que tanto creer como conocer no son predicados absolutos, en el sentido de que no se aplican directamente a oraciones que son “creídas” o “conocidas” intemporal o impersonalmente. Son predicados relativos en el sentido de que lo que es conocido es conocido “por alguien” y “en un momento” determinados. Pasar por alto este hecho puede dar lugar a muchas paradojas y extravíos. En el plano semántico, la introducción de momentos temporales (a través de una familia ordenada de índices temporales $\{t_0, t_1, t_2, \dots\}$) conlleva la introducción de una familia de situaciones epistémicas, tantas como índices temporales. Si además, queremos tener en cuenta no sólo el estado de conocimiento en el momento t , sino además aquellas situaciones epistémicas no realizadas pero que pudieron haber sido en el momento t , entonces hemos de introducir una familia de situaciones epistémicas para cada índice t_i . Esto supone la definición de conjuntos de conjuntos de conjuntos de mundos posibles. Pero si, además, tenemos en cuenta la existencia de distintos agentes epistémicos, de distintos sujetos de conocimiento (a través de una familia de agentes epistémicos $\{a_1, a_2, a_3, \dots\}$), terminamos trabajando con conjuntos de conjuntos de conjuntos de conjuntos de mundos posibles, con el alto grado de complejidad que ello supone. En la práctica, sin embargo, no es necesario casi nunca definir este “macro modelo” sino que basta con definir los elementos del modelo imprescindibles para resolver o dar cuenta del problema concreto que nos atañe.

contradicción doxástico, según el cual no podemos creer una proposición y su negación a la vez:

$$\text{PCD: } \neg(\text{Cp} \wedge \neg\text{C}\neg\text{p}).$$

Si por R entendemos la noción de creencia racional (no podemos creer lo que niegan las evidencias, ni dejar de creer lo que ya sabemos), entonces hemos de imponer una ulterior restricción sobre el marco semántico, que asegure la validez de la fórmula:

$$\text{CR: } \text{Sp} \rightarrow \text{Cp}$$

CR es válido tan pronto como imponemos el requisito de que todo mundo de la situación doxástica sea un mundo de la situación epistémica actual (esto no debe provocar la más mínima inquietud: una situación compatible tanto con mis conocimientos como con mis creencias es, claro está, una situación compatible con mis conocimientos):

$$(R.1) \text{ Para todo } d^j, d^j \subset e_a.$$

En este escueto marco podemos aprehender un primer sentido, bastante fuerte, de “credibilidad”, según el cual algo es creíble si pertenece a alguna de las creencias que son compatibles con mi estado de conocimiento actual. Expresemos mediante la letra “L” la expresión “es creíble dado mi saber actual”:

$$(D.1) \text{ Lp} \equiv \exists d^j \forall u \in d^j (u \in \#(p))$$

Una segunda noción de credibilidad, más general, aparece cuando tenemos en cuenta no sólo el estado de conocimiento “actual” e_a , sino también aquellas situaciones epistémicas $\{e_i, e_j, \dots\}$ que pudieron haberse dado en vez de ella, y para cada una de ellas tenemos en cuenta el conjunto de situaciones doxásticas correspondientes. Para ello definamos una función Π de situaciones epistémicas a conjuntos de situaciones doxásticas, tal que Π nos da para cada situación epistémica e_i , el conjunto de situaciones doxásticas $\{d^j_i, d^k_i, \dots\}$ compatibles con e_i . Π sólo ha de cumplir el requisito 1 matizado:

$$(R.1.)' \forall d^j_i \in \Pi(e_i) (d^j_i \subset e_i)$$

Simbolizamos mediante las letras “ $\Diamond C$ ” la expresión “... es creíble”. La definición a la que llegamos es la siguiente:

$$(D.2) \Diamond C p \equiv \exists e_i \exists d^j_i \in \Pi(e_i) \forall u \in d^j_i (u \in \#(p))^5.$$

Un concepto complementario al de creíble es el de “dudable”: algo es dudable si pudiera darse el caso que no lo creyera. Análogamente, definamos

⁵ Incidentalmente: merece la pena advertir que lo que es creíble en este sentido, depende de lo que es cognoscible. La dependencia aquí corre paralela a la dependencia entre lo que es creído y lo que es conocido. Si algo puede ser creído, entonces es compatible con alguno de mis posibles estados de conocimiento.

dos nociones: “E”, dudable dado el saber actual, y “ $\Diamond\text{---}C$ ”, dudable en general:

$$(D.3) Ep \equiv \exists d_i^j \in \Pi(e_a) \exists u \in d_i^j (u \in \#(\neg p))$$

$$(D.4) \Diamond\text{---}Cp \equiv \exists e_i \exists d_i^j \in \Pi(e_i) \exists u \in d_i^j (u \in \#(\neg p))$$

Por último, la oración de Austin nos lleva a la consideración de un nuevo predicado aplicado a sentencias: A, el cual equivale a “afirma que..” o “asevera que...”. Afirmar que p es, de hecho una actitud proposicional más. Una propiedad importante de la misma es que, si afirmamos que p, entonces afirmamos todas las consecuencias lógicas de p. Esta propiedad impone no sólo un importante principio rector en nuestras conversaciones sino es que la base misma de nuestras prácticas argumentativas: cuando razonamos con alguien intentamos hacerle ver cuales son las consecuencias de sus afirmaciones, consecuencias con las cuales el hablante se ha comprometido (aun cuando no fuera consciente de ellas) al aseverar lo que aseveró. Adoptemos pues el siguiente principio de afirmación de las consecuencias:

(PAC) Si Ap y q es una consecuencia lógica de p, entonces Aq.

A primera vista, no hay muchas más condiciones que A debe cumplir. Sin embargo, el ejemplo de Austin intenta mostrar que, al menos si alguien afirma sinceramente que p, entonces cree que p⁶. Claro que alguien puede mentir, y entonces no se sigue que crea lo que afirma sino todo lo contrario (el mismo Austin admite esta situación), pero entendamos por A la aseveración sincera, para evitar enredos, y establezcamos el siguiente principio de implicación de la creencia por la aseveración

(IAR) $Ap \rightarrow Cp$.

Pues bien, de todo esto, según Austin, se sigue que no es posible afirmar sinceramente que p y que no se cree que p. Pero, ¿realmente PA se sigue de las definiciones y supuestos hasta aquí adoptados? Supongamos que alguien afirma ($p \wedge \neg Cp$), por PAC se sigue que afirma p, y que afirma $\neg Cp$, y por IAR se siguen tanto Cp como $C\neg Cp$. Ahora bien, en nuestro modelo no hay nada que nos impida tener a la vez Cp y $C\neg Cp$. Puede ocurrir que p sea verdadero en todo mundo de la situación doxástica actual, y sin embargo en alguno de esos mundos no es creído que p. De ahí que el supuesto de que alguien afirme ($p \wedge \neg Cp$) no lleve a ninguna contradicción, y el principio de Austin no pueda demostrarse.

¿Quiere esto decir que el análisis formal es insuficiente? Al contrario, lo que muestra es que las presuposiciones en torno a la afirmación, la creencia y

⁶ Como se verá al final del trabajo, este principio puede ponerse en cuestión, pero en la perspectiva de Austin es irrefutable.

la sapiencia que hemos invocado no son suficientes para hacer válido TA. Ante esto, caben dos alternativas: o bien ahondar más en nuestras convenciones semánticas en busca de alguna o algunas que lo hagan válido, o bien rechazar TA sobre la base de que nuestras convenciones semánticas no son suficientes para validarlo, y buscar algún caso real que sirva como contraejemplo. A continuación ensayaremos las dos vías.

Primera Solución: creer que se cree lo que se cree

La primera solución pasa por justificar el principio de Austin bajo la adopción de intuiciones más o menos plausibles acerca de la iteración de predicados epistémicos. En “Saber y Creer” Hintikka expone una acabada teoría acerca de la iteración y combinación de operadores de sapiencia y operadores de creencia. Uno de los principios que discute es el llamado “principio de introspección”:

(P.I.) $Cp \rightarrow SCp$ ⁷

P.I. recoge la idea de que nuestras creencias nos son completamente transparentes, es decir, conocemos todas y cada una de nuestras creencias. Sobre esta base el principio de Austin es correcto:

Sup.	1	$A(p \wedge \neg Cp)$	
	2	Ap	PAC
	3	$A\neg Cp$	PAC
	4	Cp	IAR
	5	$C\neg Cp$	IAR
	6	SCp	PI
	7	CCP	CR
	8	$\neg(Cp \wedge C\neg p)$	PCC
		X	

De cualquier modo, el principio de introspección resulta bastante problemático, ¿realmente conocemos todas nuestras creencias? Dejando aparte los predicamentos del psicoanálisis o situaciones más o menos novelísticas de dobles personalidades, hay muchas ocasiones en las que tenemos serias dudas acerca de si algo forma parte o no de nuestras convicciones (¿cree en el principio de Fermat?), o lo “creemos” tan débilmente, con tan leve grado de confianza, que no sabemos si realmente lo creemos (¿cree que estará mañana a las tres en la Facultad?). Para quien tales situaciones impiden asumir el principio de introspección, una forma de salvar la tesis de Austin es recurrir a una versión débil del principio de introspección según la cual podemos no estar

⁷ Por cierto, Hintikka rechaza el principio, que es inválido en su sistema.

completamente seguros sobre nuestras creencias, pero al menos debemos tener la suficiente conciencia de ellas como para creer que las poseemos:

(P.I.D.) $Cp \rightarrow CCp$

PID es válido en nuestro modelo tan pronto como aseguramos algún tipo de “transitividad epistémica” entre mundos posibles, de tal modo que si v pertenece a la situación doxástica actual de u y w pertenece a la situación doxástica actual de u , entonces w pertenece a la situación doxástica actual de u . Hasta ahora no hemos tratado más que de las situaciones epistémicas y doxásticas actuales y posibles para el mundo actual a , pero para tratar combinaciones de predicados epistémicos debemos definir tales situaciones para cada uno de los mundos posibles que componen U . En tal marco, P.I.D es válido tan pronto como se cumple la condición referida, ya que si p es verdadero en todos los mundos pertenecientes al estado doxástico actual (sentido de Cp), entonces p es verdadero en todos los mundos que pertenecen a los estados doxásticos actuales respectivos de los mundos del estado doxástico actual del mundo actual (sentido de CCp).

Aunque esto asegura la validez de PID su validez filosófica es, en mi opinión, más comprometida. Para las creencias muy débiles no sólo es el caso que yo no sé que las tengo, sino que además no soy capaz de decidirme a afirmar si las poseo o no (“dudo” es aquí el predicado epistémico correspondiente), por lo cual no parece adecuado describir la situación diciendo que creo que creo p aunque no lo sé. Al fin y al cabo, nuestra intuición es que si es el caso que creo p , y además es el caso que p , entonces sé que p ; lo que en este caso supone que no hay mucha diferencia entre las dos versiones del principio de introspección (siendo aquí la p que creo y es el caso “Sp”).

Segunda solución: afirmar que se cree lo que se afirma

La segunda estrategia para salvar el principio de Austin se produce incidiendo en las presuposiciones conversacionales que se siguen del simple hecho de afirmar p . Para Austin, una de tales presuposiciones es que creo que p , pero, como vemos, esto no es suficiente. Podemos considerar algo más fuerte: que de afirmar p no sólo se sigue que creo p , sino que además *al afirmar p afirmo que creo que p* . Esto son dos cosas muy distintas. La primera era recogida por AR, la segunda se enuncia así:

(IAR2) $Ap \rightarrow ARp$.

Es fácil ver que AR2 (junto con AR) implica el principio de Austin:

Sup. 1 $A(p \wedge \neg Rp)$

2 Ap PAC

3 $A\neg Cp$ PAC

4 Cp	IAR
5 $C \neg Cp$	IAR
6 ACp	IAR2
7 CCp	IAR
8 $\neg(Cp \wedge C \neg p)$	PCC
X	

No sé hasta qué punto los filósofos del lenguaje o los lingüistas aceptarían las consecuencias que se siguen de este análisis. ¿ Realmente “que creo que p” es parte de lo que quiero decir al afirmar “p” a secas? ¿No haría eso inútil la misma existencia de la palabra creer, así como redundantes todas las oraciones que comienzan por “yo creo que...” o algo parecido? Además, y esto tal vez sea lo más problemático, IAR2 conduce a una serie de infinita de afirmaciones incluidas en mi simple aseveración de que p. En efecto, por IAR2, si afirmo p se sigue que afirmo ACp, pero de ahí, por IAR2 otra vez, se sigue que afirmo ACCp, y además que afirmo ACCCp, ACCCCp... Todo esto problematiza IAR2, aunque de por sí tal vez no sea conclusivo para su rechazo.

Hasta aquí con lo que respecta al primer miembro de TA, pero ¿qué ocurre con el test de Austin para la “implicación no ordinaria”? En nuestra versión formal de la tesis de Austin, establecíamos que podía darse para alguna proposición p, que afirmemos que p pero que sea posible no creer que p (sea posible dudar que p). Tal como hemos definido las situaciones doxásticas es fácil ver que este caso no es, en absoluto, aberrante en nuestro modelo: corresponde a una situación en la que p es verdadero en todo mundo de la situación doxástica actual pero es falso en algún mundo de alguna de las situaciones doxásticas posibles (como recoge la condición D4). Lo cual nos permite dar un primer “jaque” a Austin: es posible validar la primera parte de su tesis sin que ello nos aboque a la negación de la segunda.

Tercera solución: negar TA

Las consideraciones nos hacen ver que el Principio de Austin se asienta sobre presupuestos problemáticos (aunque ambos han sido aceptados por gran numero de investigadores; tan o más grande que el número de los que lo han rechazado). Cabe plantearse, en consecuencia, si no existirán situaciones (aunque inusuales o tal vez remotas) en las cuales alguien pueda afirmar que hay un gato en la alfombra pero que no cree tal cosa en absoluto. Al fin y al cabo, que una afirmación cause extrañeza no significa necesariamente que haya una contradicción: si alguien afirma, por ejemplo, que mi corbata es azul

mientras dirige sus ojos directamente a mi impoluta corbata de seda roja, consideraría la afirmación ciertamente extraña. Pero antes de concluir que mi interlocutor está violando alguna de “las convenciones semánticas que rigen el uso de las palabras” procuraría comprobar si no concurre algún hecho excepcional y muy poco frecuente que hace que su afirmación tome sentido: intentaría averiguar, por ejemplo, si el individuo es daltónico, o si tiene algún desorden en su conceptualización de los colores.

Análogamente, es posible encontrar ejemplos en la literatura de la psicología clínica en la cual individuos con problemas para obtener una conciencia del “yo” o para acceder al mundo de las creencias propias estarían precisamente en la situación en la que afirmar que hay un gato sobre la alfombra pero que no lo creen es lo que mejor refleja el estado de cosas real.

Es más, tal vez ni siquiera haga falta recurrir a caso tan “anómalos” (me atrevería a decir “enfermos”, si no temiera a las connotaciones negativas). Al fin y al cabo, la afirmación por sí sola no tiene porque involucrar la creencia. Afirmar que *p* es comprometerse con la verdad de *p* y las consecuencias que pueda tener *p*, y esto puede darse incluso aunque no tengamos auténtica fe sobre la exactitud de *p* (aunque esto nos lleve a poner en cuestión IAR). Generalmente paso a aseverar una cosa en el momento en que creo en ella, pero también puedo comprometerme con una teoría científica, un ideal social o una descripción de qué-fue-lo-que-paso por otros motivos: simpatía o confianza respecto a los que lo defiende, intereses personales (tal vez económicos), una intuición que me dispongo a seguir, o simple coerción. Al fin y al cabo: ¿no afirmo Galileo que la tierra era el centro inmóvil del Universo ante el tribunal de la Iglesia?. Alguien puede opinar que mentía; sin embargo, la Iglesia no pedía a Galileo que **creyera** que la tierra estaba suspendida en el centro del espacio, sino tan sólo que lo **aseverara**, y en este sentido Galileo no fue deshonesto al hacer su promesa (aunque sí, tal vez, un poco cobarde). Para quienes así lo vemos, al añadir la famosa coletilla “y sin embargo se mueve” (si es que en realidad fue dicha) el famoso físico pisano no se contradijo, sino tan sólo introdujo una importante (y no sutil) matización.

BIBLIOGRAFÍA

- Austin, John L. (1961): “The Meaning of a Word”, *Philosophical Papers*, Oxford U. P. 1961. (Se cita según la traducción española *Ensayos Filosóficos*, Alianza Editorial, Madrid 1989).
- Hintikka, Jaako (1962): *Knowledge and Belief*, Ithaca, Cornell University Press, Nueva York
- Martínez, Concepción; Sagüillo, José Miguel; y Vilanova, Javier (1997): “Fitch’s Problem and the Knowability Paradox: Logical and Philosophical Remarks”. *Logica Trianguli* 1, Nantes 1997, 1-22.
- Vilanova Arias, Javier (1998): “Nociones epistémicas modales”, transcripción de la comunicación presentada al II Congreso de la Sociedad Española de Filosofía Analítica,

Madrid.

Vilanova Arias, Javier (1999): "Un análisis de la noción de cognoscibilidad desde la semántica de mundos posibles", *Theoria* vol. 14, nº 36, págs. 413-429.

Javier Vilanova Arias

Universidad Complutense de Madrid

e-mail: vilanova@eucmax.sim.ucm.es



Ontology

Ontología





EL REALISMO METAFÍSICO Y LA NECESIDAD DE RENOVAR LA FILOSOFÍA*

MANUEL BARES PARTAL
Universidad Autónoma de Barcelona

ABSTRACT: Putnam's recent works insists in what he calls "the problem of Realism". The question of realism is divided in two main options. On the one hand the realist claim for an absolute conception of reality in which human perspective has no place. In the other, for relativism that perspective is the only realist option in metaphysics. In Putnam's view, Pragmatism is the option that conserves realism and, at the same time, solves the genuine relativistic point. Nevertheless, Putnam's Pragmatism is an incomplete theory. Wittgenstein's "following a rule" is presented in this work as an option to solve problems that are still open in Putnam's books.

KEYWORDS: Metaphysical realism, relativism, pragmatism, language game, fact \ value, following a rule.

De un tiempo a esta parte, los trabajos de Hilary Putnam vienen planteándonos un reto difícil de comprender, cuanto más de acometer con garantías. Ni más ni menos que el reto de "renovar la filosofía". Según este autor, la filosofía hace tiempo que se encuentra prisionera entre las dos puertas de una estancia herméticamente sellada. En una de ellas vigila la Escila del realismo, tras un mundo de objetos que podemos ver y nombrar, pero no se nos permite el acceso a lo que se encuentra más allá de la puerta. En la otra, el Caribdis del relativismo vigila que todo lo que llegue a nosotros tenga la inevitable debilidad de la apariencia¹.

El reto de renovar la filosofía es planteado por Putnam como el intento de superar la ilusión de necesidad de atravesar una de estas dos puertas rechazando la otra. La intuición que intenta compartir con nosotros es la de que en cierta forma, ambas puertas no conducen a estancias distintas, sino que ambas nos llevan a una misma gran sala, la del realismo metafísico que, tras

* Este trabajo ha sido financiado en parte por el proyecto de MEC. DGES. PB97-0192.

¹ Putnam (1990 y 1992)

provocar la tensión interna que acabamos de denunciar, parece alejarnos de la irreprochable objetividad del realismo, merced a las dudas relativistas al tiempo que deja en suspenso la sensatez de un grado moderado de relativismo bajo la sospecha de una arbitrariedad irresponsable.

En lo que sigue vamos a tratar de estudiar un poco más detenidamente esta cuestión. Soy consciente de que un estudio detallado de las presuntas debilidades del realismo metafísico así como de, en los términos wittgenstenianos en que Putnam se expresa habitualmente, aplicar la terapia adecuada al problema que la filosofía debe sanar, excedería con mucho las pretensiones de una exposición como esta. Por esta razón nos limitaremos a tratar de hacer un poco más claros los puntos centrales que marcan la cuestión. En la primera sección trataremos de reconstruir, de manera breve, el planteamiento del problema tal y como Putnam lo ha ido conformando a lo largo de sus obras². Tras esta reconstrucción, que será más una exposición del punto de vista de Putnam que un análisis de sus argumentos, trataremos de entrever el camino que Putnam nos invita a seguir como posible vía de superación de la incomoda dualidad. Sin embargo, la terapia de Putnam no siempre está a la altura de su diagnóstico³. Tras la exposición de los supuestos pragmatistas que Putnam nos invita a compartir, la segunda sección acabará con algunas dudas sobre la manera de entender su enfoque pragmatista. Y en la última sección señalaremos la posibilidad que la filosofía del segundo Wittgenstein nos abre para tratar de superar el problema planteado por Putnam sin necesidad de abandonar necesariamente el seguro realismo, ni el sensato relativismo.

1. El planteamiento del dilema

Es difícil entender lo metafísicamente real como algo distinto a aquello que está en el mundo y que justifica nuestras percepciones. Esta mesa, por ejemplo, es real sólo en el sentido de que es un objeto situado en el espacio descriptible a través de propiedades primarias. Que es un objeto práctico, que es de tal color o que es dura, lisa y fría no son propiedades que pertenezcan al objeto como tal, sino que más bien son maneras en las que nosotros, los seres humanos, la percibimos.

El problema es que sólo podemos percibir un objeto de esta forma. Cuando percibimos la mesa en cuestión, lo que tenemos es la sensación de una superficie dura, lisa y de tal color, y nada más. Sin embargo, cuando lejos de la necesidad de apoyar algo en la dura y coloreada superficie de la mesa, nos preguntamos qué es realmente el objeto que hay delante de mí, cómo es al margen de mi manera de verlo, cómo es *en realidad*, no buscamos la respuesta en nuestras necesidades idiosincrásicas sino en la rotunda autoridad

² Especialmente en Putnam (1992, 1994 y 1995)

³ Ver Putnam (1995)

científica. El objeto que tengo ante mí es un compuesto de átomos y vacío individualizado en virtud de su posición en el espacio respecto a otros objetos.

A primera vista, diremos que la primera explicación no es una verdadera explicación de lo real, sino una explicación que depende de una perspectiva cuyo foco son nuestras particularidades. Particularidades que pueden ser físicas, psicológico-cognitivas o culturales. La segunda explicación, en cambio, responde a un esquema científico, se acerca más a lo real. La virtud principal de esta segunda forma de concebir lo real es que podemos, atendiendo a las propiedades de los átomos y del vacío en conjunción con las propiedades de nuestra fisiología y psicología, dar una explicación de la dureza, frialdad y colorido de la mesa que percibimos. El objeto científico es explicativamente anterior al objeto percibido, y esto le otorga un estatus ontológico mayor. La distinción de Sellars⁴ entre una imagen manifiesta del mundo frente y una imagen científica del mismo, así como las tesis de B. Williams⁵ acerca de una concepción absoluta de lo real frente a una perspectiva determinada de la misma son teorías que se basan en la sencilla intuición que estamos tratando de explicar.

No obstante, hasta aquí la imagen es incompleta. Tenemos un mundo formado por entes reales e inalcanzables que de alguna manera determinan nuestra particular forma de ver la realidad. Falta explicar cómo. Se me ocurren tres maneras de enfrentarse a este problema:⁶

1. La vinculación de la perspectiva con la realidad es una cuestión de necesidad explicativa. Cuando tratamos de comprender mejor qué son los objetos que percibimos, siempre encontramos, al final de nuestras razones, la necesidad de postulación de una realidad criterio que sostenga la explicación. La imagen es similar a la que la distinción moderna entre cualidades primarias y secundarias sugiere. Las primarias forman el esquema básico sobre el que se apoyan las secundarias. Este es el punto de vista de autores como B. Williams, por ejemplo.
2. Los objetos que percibimos son fruto de la eficacia causal de lo real. Los objetos reales causan representaciones en nuestras mentes. Estas representaciones, o sense-data, conforman nuestra realidad que, si bien no son lo real mismo, son fiables en tanto que su modelo lo es. El criterio de adecuación de la representación a lo real puede ser de primera persona, como en el empirismo clásico, o de tercera, como en las actuales teorías de la filosofía de la mente, pero ambas tienen en común la necesidad de postulación de algún tipo de imagen mental que funcione como representación.

⁴ Sellars (1953).

⁵ Williams (1978).

⁶ En Putnam (1992 y 1995).

Es importante tener en cuenta que, aunque 1 y 2 no son incompatibles, su vinculación tampoco es estrictamente necesaria. Puede defenderse 1 sin postular sense- data, como en el caso de Williams, y se puede ser un defensor de las imágenes mentales al tiempo que se mantiene un moderado escepticismo respecto al alcance ontológico del criterio explicativo.

3. La tercera manera de enfrentarse a cómo se vinculan realidad y perspectiva humana es negando, bien que dicha vinculación se dé, bien que podamos tener acceso epistémico a ésta. Desde el relativismo, que encarna esta tercera manera de entender la relación, se establecen convenciones que asumen la misma función que una concepción absoluta de la realidad. No hay un criterio metafísico más allá de nuestras propias convenciones y que justifique nuestro conocimiento de la realidad, si bien estas cumplen, para el relativista, el papel que lo real habría de cumplir si existiera.

Estas tres formas de entender la relación pueden, en realidad, resumirse en dos. Las dos primeras conformarían la hipótesis del realismo. Nuestras percepciones y los enunciados de nuestro lenguaje poseen un contenido que aparece determinado por un mundo absoluto que es el único real y que funciona como el único criterio válido de corrección. Por otra parte, la tercera opción es la del relativismo extremo. Todos los puntos de vista son igualmente verdaderos o falsos ya que no hay, fuera de las convenciones, un criterio que nos permita establecer verdades universales. La comunicación entre comunidades que se atienen a diferentes convenciones es mera apariencia. La pregunta es: ¿por cuál de las dos opciones hemos de decidarnos?

El planteamiento mismo de la cuestión no parece dejar lugar a dudas, o realismo o relativismo, o mundo absoluto o convenciones. Por una parte, la opción realista nos ofrece la irrechazable posibilidad de comprender el mundo tal y como es. El mero hecho de ser la posición que mejor se adapta a la investigación científica es ya suficiente aval como para sentir simpatías por esta opción. La promesa de llegar a conocer verdades universales es lo suficientemente tentadora para el filósofo como para hacer de esta opción la propia del sentido común.

De todas formas no podemos rechazar el relativismo con tanta facilidad. La hipótesis relativista puede ponerse bajo sospechas, pero deja siempre un cierto resquemor. Posee, como dice Putnam, “esa cierta verdad que esconde un sano relativismo”. Por otra parte, si el realismo es la opción que mejor se adapta a las exigencias de la ciencia, el relativismo es a su vez la que mejor da cuenta de las relaciones humanas.

Ésta es la cuestión con que se enfrenta Putnam. ¿Cómo tratar de entender el mundo de manera que no tengamos que prescindir ni del necesario

realismo ni del útil relativismo?⁷. Para Putnam la cuestión es compleja. La respuesta a este interrogante hay que buscarla en aquello que tienen en común toda una serie de autores a lo largo de la historia de la filosofía. Autores como Kant, William James, Dewey o Wittgenstein. Todos ellos poseen, para Putnam, los datos del punto de partida para el nuevo camino metafísico que ha de seguir la filosofía. Todos ellos son, en mayor o menor medida, filósofos pragmatistas.

2. El realismo pragmático de H. Putnam

Uno de los síntomas de enfermedad más evidentes que Putnam encuentra en el realismo metafísico es el de separar radicalmente entre hecho y valor. Una concepción que pueda dar cuenta de aquello que el realismo y el relativismo pueden aportar a la filosofía ha de situarse necesariamente en un marco donde lo real no suponga la exclusión de lo que el sujeto aporta a la determinación de la realidad, al tiempo que se respeta la necesaria objetividad que requiere la ciencia. Sólo a través de la interrelación entre hecho y valor es posible, para Putnam, superar el dilema que la filosofía arrastra desde la modernidad.

En este sentido, Putnam apuesta decididamente por aceptar el criterio de la acción que el pragmatismo clásico oponía, de manera similar, al positivismo y al idealismo que entonces representaban los cuernos del dilema. Las tesis pragmatistas de William James y su teoría de la verdad como satisfacción son, para Putnam un buen ejemplo de cómo hecho y valor se funden para dar forma a lo real⁸. Verdadero es aquello que resulta satisfactorio para los intereses del sujeto. Éste no trata de entender el mundo desde fuera, sino que de alguna manera se enfrenta a la realidad desde su capacidad de vivir en el mundo, de interactuar con él a través de sus acciones racionales o prácticas⁹.

De esta manera, tratar de entender el mundo al margen de su interrelación con él es algo que no tiene sentido. Putnam rechaza, basándose en este comienzo, la idea de un mundo consistente en una cantidad de objetos fijados de una vez al margen de nuestra capacidad de conocerlos. El dominio de lo real es un dominio abierto. En este sentido, Putnam retoma también otra de las propiedades de la filosofía pragmatista. El falibilismo que se encuentra, entre otras en las tesis de Charles S. Peirce¹⁰ parece ser la única posición epistémica que pueda responder al reto a la vez conformador y espectador del sujeto.

⁷ Ver Putnam (1994 y 1995).

⁸ Putnam (1994 y 1995).

⁹ Ver James (1979).

¹⁰ Peirce (1934).

Aparentemente este enfoque pragmatista tiene que afrontar algunos problemas: a) ¿no se confunde aquí verdad y confirmación de creencias? b) ¿no es el pragmatismo una nueva manera de expresar una posición fenomenalista? c) ¿no suponen las tesis de Peirce una manera particular de entender el positivismo imperante a finales del diecinueve?¹¹ Sin embargo cada una de estas cuestiones sólo tienen sentido si partimos de la imagen que surge del realismo metafísico.

- a) Sólo si se piensa que la cuestión de decidir si algo es verdadero depende de la adecuación de lo real a nuestras representaciones puede acusarse al pragmatismo de confundir verdad y confirmación, ya que para el pragmatismo es a través de la confirmación de creencias como el sujeto determina lo real. Si la propia acción del sujeto supone un marco válido para la determinación de lo que entendemos por real, entonces no puede separarse lo real de la acción, el hecho del valor, la verdad de la confirmación de las creencias. El pragmatista no confunde una cosa con la otra, es el realista metafísico el que se empeña en separar dos cosas que no son más que una y la misma.
- b) El fenomenalismo es sólo una tesis extrema del realismo metafísico. No niega que algo cause nuestras percepciones, sólo que podamos ir más allá de éstas. Es llevar al extremo la intuición que indicábamos al principio, existe una realidad que causa nuestras percepciones, pero a la que no tenemos acceso más que a través de nuestra manera de percibirla. No podemos separar el mundo de nuestra manera de percibirlo. El pragmatismo no entiende que nuestras acciones nos oscurezcan el acceso al mundo que las causa, sino que parte de la idea de que éstas forman parte del mismo, y no son ni separables ni excluyentes.
- c) El positivismo cree que el mundo es el que es, y es nuestro conocimiento el que avanza de forma lineal. Para el pragmatismo, en tanto que nuestro conocimiento forma parte del mundo, este se va constituyendo a medida que nuestro conocimiento va avanzando. Sólo muy superficialmente puede acusarse al filósofo pragmatista de positivista.

Esto es lo que Putnam recoge del pragmatismo clásico. La interrelación hecho valor que en Putnam encarna una manera de expresar las tesis del holismo. En efecto, hasta aquí hemos contestado a una cuestión ¿qué es lo real? Falta por contestar a otra ¿cómo se conforma lo real en relación a la acción? Para Putnam, la relación constitutiva hecho valor es holista. Los hechos presuponen el valor a través de un marco de interpretación determinado, y dicho marco no es, a su vez, algo aparte del que conforma la

¹¹ Putnam (1994 y 1995).

interrelación hecho valor¹². En este sentido, Putnam recoge lo que le parece interesante de su interpretación del pragmatismo, su pluralismo.

Aun podría argumentarse, contra Putnam, que es necesario acudir al realismo metafísico como criterio último de vinculación entre los distintos marcos de interpretación. Al fin y al cabo lo que todos ellos tienen en común es la idea de que todos son formas de entender lo real. A esto Putnam responde con la tesis wittgensteniana de los “aires de familia”¹³. Lo que tienen en común las distintas interpretaciones no es una única realidad que las vincule sino, una vez más, el holismo que subyace al concepto de “parecido de familia”.

Sin embargo hay algo incompleto aquí. El pragmatismo nunca fue una tesis completa sobre metafísica. Los autores pragmatistas apoyaban sus bases sobre distintas teorías acerca de cómo se conformaba la realidad entorno al holismo, como lo llama Putnam, que forman los conceptos de hecho y valor. A esta cuestión Putnam sólo parece responder de forma negativa. No al “radical empiricism” de James, no al complejo conceptual de Mc. Dowell, no a las tesis sobre la causalidad indirecta de Davidson...¹⁴ La cuestión de que hecho y valor se vinculan formando una concepción holística de la realidad esta abordada en diversos lugares, pero ¿bajo qué criterios se rigen hecho y valor para ser constituyentes de lo real? La siguiente sección supone una tímida tentativa de respuesta bajo la influencia de la filosofía de Wittgenstein.

3. El pragmatismo de Wittgenstein como alternativa al realismo metafísico

A todo lo largo de este trabajo he intentado explicar por qué Putnam defiende que la pregunta acerca de lo que hay más allá de nuestra manera de ver el mundo es una falsa pregunta. Parece exigir algún tipo de respuesta por parte del filósofo o del científico pero estos, por lo implícito en la pregunta misma, no pueden ir más allá de su manera de ver el mundo. La pregunta no tiene respuesta.

La manera en la nos relacionamos con el mundo es muy compleja. Si miramos cualquiera de los ejemplos de juegos de lenguaje que Wittgenstein inventa y analiza en *Los Cuadernos Azul y Marrón* o en las *Investigaciones Filosóficas*, se comparta o no su filosofía, uno tiene que aceptar que quizás tenga razón en una cosa: es muy complicado dar cuenta de la marea de formas en las usamos las palabras de nuestro lenguaje. Tratar de resumirlas todas en una única teoría del significado es una labor que, cuando menos, no resultará fácil.

¹² Putnam (1995)

¹³ Putnam 1992 y 1995).

¹⁴ Putnam (1995).

Lo que yo propongo en este trabajo es una reflexión sobre la complejidad de nuestras relaciones con el mundo, y sobre la aparente ingenuidad de tratar de reducirlas al único criterio del realismo metafísico. La tesis de un mundo absoluto trata de explicar, entre otras cosas, la naturaleza de nuestra imagen de la realidad, pero si vamos un poco más allá parece que debemos completar el círculo y aceptar que nuestras necesidades y relaciones son también constitutivas de lo real.

El realismo metafísico nos seduce por la simplicidad de su planteamiento. Las cosas son como son porque el mundo oculto de lo real así lo impone, y punto. Sólo hay una forma de explicación de lo real.

Sin embargo, la realidad no se constituye de una sola manera. Los objetos de la ciencia se constituyen a través de un modelo explicativo y predictivo mientras que los objetos que llamamos mesa, ordenador etc. se constituyen merced al uso que hacemos de ellos. ¿Por qué son más reales los primeros que los segundos?

La respuesta que hemos considerado en las primeras partes de este trabajo consistía en que el orden explicativo debía ser el orden ontológico puesto que parece encontrarse en un estadio anterior. Para finalizar este trabajo pido un experimento mental. Supongamos que encontramos a alguien con un criterio práctico excelente pero con un desconocimiento y un escepticismo sobre la ciencia casi totales. La literatura y la imaginación abundan en este tipo de sabios populares. Imaginemos que sólo cree lo que ve y que la explicación de que los colores que ve no son más que disposiciones de una superficie frente a la luz le resulta increíble. Podemos enseñarle que si entendemos así los colores, podemos construir ecuaciones que nos resulten útiles a la hora de saber qué ocurrirá más tarde con esos colores. Nuestro sabio popular acepta estas ecuaciones pero sólo porque le resultan útiles, y desde luego sólo porque ve, y está convencido de que ve, colores. Si ahora le decimos que la ciencia de los colores no es sólo una herramienta para nuestra comprensión de lo real, sino la explicación metafísica de lo real mismo, posiblemente nuestro personaje se sienta estafado. ¿pero, si ya no hay colores, para qué quiero yo algo que los explique? Diremos que no es esta la forma adecuada de entender la ciencia y en cierta medida será cierto, pero hay algo más.

El sabio popular no sigue las reglas que rigen la práctica científica, sólo aplica términos de ésta a su propia manera de entender la realidad. Desde este punto de vista no entiende que, en última instancia, la mesa que tengo ante mí no sea más que un conjunto de partículas elementales y vacío. Su perplejidad se explica porque trata de exigir a los objetos de la ciencia los mismos criterios de realidad que exige a su particular manera de entender el mundo a través de unos criterios más prácticos que teóricos. Aceptamos esto y decimos que el sabio popular no entiende el tipo de pretensiones que

achacamos a la práctica científica, en otras palabras, no se atiene a las reglas que rigen el juego de lenguaje de la ciencia.

El realista metafísico debe estar de acuerdo con esto, pero no va más allá. Sin embargo el análisis del sabio popular es distinto. Para éste, es el científico el que no se atiene a su criterio de realidad, el que no sigue las reglas correctas. Decir que las incógnitas de las ecuaciones de la física son más reales que la lluvia y el calor que hacen crecer la cosecha es incomprensible. De alguna manera nuestro sabio es también un realista metafísico en tanto que piensa que sólo hay una única manera de entender el mundo y las distintas formas de relacionarnos con éste. Es fácil seguir imaginando el diálogo entre un científico idealmente entusiasta y nuestro sabio popular, pero sólo nos llevaría a una conclusión: no se ponen de acuerdo no porque uno de ellos esté en posesión de la “Verdad” con mayúsculas y el otro no, sino porque están hablando de cosas distintas. Están usando criterios diferentes que se atienen a reglas diferentes.

Williams llamaba a esto “perspectivas”¹⁵, dando por supuesto que es posible prever una forma universal de describir el mundo. Pero en tanto que esa posible forma universal de descripción se atiene a determinadas reglas, sólo podemos decir que, aun si fuera posible, presentaría la paradoja de ser una perspectiva más. Una de las múltiples imágenes que nos ayudan a componer el complejo entramado que conforma la realidad.

El relativista extremo, sin embargo, aunque acepta estas conclusiones, participa en el fondo de la misma ilusión del defensor de la absolutividad¹⁶. Sólo es concebible una única realidad. Si no es posible acceder a ella, entonces nuestros discursos son todos igualmente falaces, y sólo las convenciones pueden guiar nuestra investigación. Sin embargo uno no puede tomarse esto seriamente. El pesimismo del relativista sólo es parte de la nostalgia de la tesis rechazada. No tiene sentido la idea de una imagen universal de lo real que sirva para nivelar nuestras propias perspectivas bajo el único rasero de lo falaz. El relativismo extremo es un sinsentido que niega lo que afirma.

Puede objetarse que mi dibujo de la situación no es del todo justo. No puedo pretender convencer a nadie de que el cientificismo tienda a anular la dimensión más práctica del conocimiento o que el relativista sea tan extremo que piense que la ciencia no dice nada más del mundo de lo que ya dice el sabio popular. Muy pocos llegan a estos extremos. Sin embargo es precisamente esta necesaria moderación de las tesis la que desmiente la necesidad de un modelo rígido y extremo de determinación metafísica.

Nos relacionamos con el mundo a través de nuestras acciones, y estas son de tantas maneras como permiten las reglas que seguimos para llevarlas a

¹⁵ Williams (1978).

¹⁶ Putnam (1990 y 1992).

cabo. Primar unas en detrimento de otras sólo puede hacerse ateniéndose a ciertas reglas, y esto supone que se hace ya desde dentro de una de ellas.

El científico y el relativista no son contrarios ni están totalmente equivocados en su concepción del mundo. Simplemente atienden a distintos intereses en su búsqueda de un modelo de realidad, dice Putnam, también se puede decir que se atienen a distintas reglas.

La noción wittgensteniana de seguir una regla¹⁷ tiene los elementos que necesitamos para dar cuenta de la relación entre hecho y valor. Es un criterio que se fija desde lo objetivo y que tiene en cuenta al sujeto. Las reglas se fijan atendiendo a las necesidades del sujeto “en el mundo”, de manera que es ahí donde hecho y valor conforman una unidad. Todo juego, toda forma de relación con el mundo exige una acción que ha de atenerse a ciertas reglas, por lo que la pregunta de cómo es el mundo sin estas reglas no tiene sentido, porque ella misma esta sujeta a las reglas que le impone la acción epistémica. La diferencia con el relativismo es que este enfoque no supone que no haya una verdad, o que esta se encuentre únicamente en las convenciones humanas¹⁸. El relativismo se esfuerza en contestar, aunque sea negativamente, a una pregunta que no tiene respuesta. Además, las reglas no son convenciones, sino fruto de la interrelación entre hecho y valor, de la acción del hombre en el mundo, de la realidad.

No puede negarse que nuestras acciones se encuentran estrechamente vinculadas a nuestra manera de entender el mundo. De esta manera, William James estaba en lo cierto cuando afirmaba que la verdad era satisfacción al tiempo que rechazaba las críticas que se cebaban en contra de un posible rechazo de cualquier realidad ajena a la psicología del sujeto. Sin embargo James vuelve a caer en la trampa de intentar explicar a través de un único tipo de relación sujeto-objeto las múltiples formas de satisfacción. El “empirismo radical” reafirma la imagen que su principio de satisfacción parecía eficazmente superar.

De esta manera las conclusiones de Putnam vuelven a reafirmarse al final del trabajo. La filosofía, así como la ciencia, han de renovarse. Pero esta renovación no debe afectar a sus métodos ni a sus resultados, sino a la manera en la que éstos influyen en la forma de entendernos a nosotros mismos y al mundo. La investigación filosófica no tiene que moverse entorpecida por el dogmatismo de los realistas extremos o de los relativistas, sino que debe guiarse por la necesidad de comprender toda la incuestionable complejidad del hombre y del mundo. Simplificar esta complejidad y cerrar los ojos al resto de cuestiones no es explicar nada. La filosofía que Putnam exige ha de hacernos comprensibles las distintas maneras de entender lo real, incluida la

¹⁷ Wittgenstein (1953).

¹⁸ Ver Wittgenstein (1969).

científica. Ha de poner a todas estas maneras un rostro reconocible, un rostro también humano.

BIBLIOGRAFÍA

- Fogelin, R. J. *Wittgenstein*. Routledge and Kegan Paul, London, 1976.
- Hookway, C. *Peirce*. Routledge and Kegan Paul.. London, 1985.
- James, W. *Pragmatism and the meaning of truth*. Harvard University Press, Cambridge. Massachusetts, ed. 1979.
- James, W. *Essays on Radical Empiricism*, Jarvard University Press. Cambridge, Massachusetts, ed. 1976.
- Kenny, A. *Wittgenstein*. Allen Lane, London, 1973.
- Pears, D. *The False Prison*. NY Clarendon, Oxford, 1988.
- Peirce, Ch. S. *The Fixation of a Belief*. En *Collected Papers*. Harvard University Press. Cambridge, Massachusetts, 1934.
- Putnam, H. *Reason Truth and History*. Cambridge University Press, Cambridge/New York, 1981.
- Putnam, H. *The Many Faces of Realism*. Open Court, La Salle, Illinois, 1987.
- Putnam, H. *Representation and Reality*. M.I.T. Press. Cambridge, Massachusetts, 1988.
- Putnam, H. *Realism with a Human Face*. James Conant (ed.). Harvard University Press, Cambridge, Massachusetts, 1990.
- Putnam, H. *Renewing Philosophy*. Harvard University Press, Cambridge, Massachusetts, 1992.
- Putnam, H. "Sense, nonsense and the senses. An Inquiry on the Powers of Human Mind". *Journal of Philosophy*, New York, 1994.
- Putnam, H. *Pragmatism*. Blackwell, Oxford / Cambridge, Massachussets, 1995.
- Quine, W.V.O. "Two Dogmas of Empiricism." En *From a Logical Point of View*. Harvard University Press, Cambridge, Massachusetts, 1961.
- Williams, B. *Descartes. The Project of a Pure Inquiry*, Penguin Books, London, 1978.
- Wittgenstein, L. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953.
- Wittgenstein, L. *The Blue and Brown Books*. Basil Blackwell, Oxford, 1958.
- Wittgenstein, L. *On Certainty*. Basil Blackwell, Oxford, 1969.

Manuel Bares Partal.

Universidad Autónoma de Barcelona.

e- mail: mbares@seneca.uab.es



THERE IS NO TRUTH - THERE MIGHT BE NOTHING **Elenchos of Meta-Theoretical Negative Statements** **and Ontological Factualism**

FRANCA D'AGOSTINI
University of Turin

ABSTRACT: Meta-theoretical negative statements, like “there is no truth” are easily refuted by the Aristotelian procedure called *elenchos* (see Met. IV), that is by noticing that the rule they express cannot be applied to them: if even “there is no truth” (from now on: TNT) is false, some truth must exist. This classical device has been recently activated in different contexts and by different authors: for instance by Thomas Nagel in his criticism of relativism on the plan of an analysis of transcendental conditions of reason (see: 1997).

But - as a formal analysis shows: see Galvan, “A Formalisation of Elenctic Argumentation”, 1995 - without the premise of the necessary existence of things (or some other ontological commitment) *elenchos* is inconcluded (it turns into regressus ad infinitum). Hence there is a positive connection between *elenchos* as an argumentative tool and the ontological stance called *factualism*, or *metaphysical realism*: namely, the *elenchos* of truth holds iff the possibility of “nothing” as a state of affairs (the eventuality expressed by “there might be nothing” = TMN) is previously denied; or iff the existence of things is asserted beyond any doubt. So $\neg$ TNT is necessarily true iff also $\neg$ TMN is true.

This last consideration admittedly justify epistemological and ontological *nihilism* (TNT + TMN). The aim of the paper is showing how all that can be, on the contrary, the starting point of a kind of ontological realism. Namely: some form of trans-ontological realism (or trans-platonism) which admits the alternative recurrence of *elenchos*, joined to the modal precariousness of things, and both as *theoretical* “states of affairs”, states of affairs belonging to the realm of theory.

KEYWORDS: Ontology, history of logic, truth, *elenchos*, *factualism*, *nihilism*.

Introduction

Meta-theoretical negative statements, like “there is no truth” are easily refuted by the Aristotelian procedure called *elenchos* (see Met. IV), that is by noticing that the rule they express cannot be applied to them: if even “there is no truth” (from now on: TNT) is false, some truth must exist. This classical device has been recently activated in different contexts and by different

authors: for instance by Thomas Nagel in his criticism of relativism on the plan of an analysis of transcendental conditions of reason (see: 1997).

But - as a formal analysis shows: see Galvan, "A Formalisation of Elenctic Argumentation", 1995 - without the premise of the necessary existence of things (or some other ontological commitment) elenchos is inconcluded (it turns into regressus ad infinitum). Hence there is a positive connection between elenchos as an argumentative tool and the ontological stance called factualism, or metaphysical realism: elenchos holds iff the possibility of "nothing" as a state of affairs (the eventuality expressed by "there might be nothing" = TMN) is previously denied; or iff the existence of things is asserted beyond any doubt. So $\neg$ TNT is necessarily true iff also $\neg$ TMN is true.

This last consideration admittedly justify epistemological and ontological *nihilism* (TNT + TMN). The aim of the paper is showing how all that can be, on the contrary, the starting point of a kind of ontological realism. Namely: some form of trans-ontological realism (or trans-platonism) which admits the alternative recurrence of elenchos, joined to the modal precariousness of things, and both as *theoretical* "states of affairs", states of affairs belonging to the realm of theory.

1. Is it true that there is no truth?

Apparently, there is no problem concerning TNT: it is a simple self-refuting statement. As St. Augustin first explained (Soliloquia, II, 1, 1): if truth does not exist, hence *it is true that TNT* (TTNT), so truth necessarily exists. This is an application of the so-called equivalence thesis (if "p" then "p is true"), but with an anti-deflationist conclusion: truth is exalted, not deflated, even if the premise is the same: the ubiquity of truth.

From the ubiquity of truth St. Augustin demonstrated the immortality of soul and the existence of God. Truth is *eternal*: if the world finishes, it will be true that the world has finished; truth is *undeniable*, as above seen (if TNT, then TTNT): hence it is *necessary*, and *unavoidable*. But God is the eternal and necessary truth, and the soul is the place in which truth takes place, so God must undeniably exist, and the soul must be eternal.

The equivalence $TNT \equiv TTNT$ is probably the weakest point of the argument, but it has been accepted by most commentators. Even St. Thomas Aquinas¹, who criticised Augustin's metaphysical conclusions (i. e. the immortality of the soul and the necessary existence of God), accepted it, and agreed on the self-refuting nature of TNT. Some commentators on the contrary (for a time also the Augustinian St. Bonaventure²) considered the equivalence fundamentally wrong. John Duns Scotus observed that "truth

¹ *Summa Theologiae*, I, 2, 1, ob. 3.

² See Charron-Doyle: 1996, 249

does not exist” can only be translated into “‘truth exists’ is not true”. Buridan, who probably gave the most “modern” analysis of the problem, noted that “homo est animal” does not mean “homo est animal is true”, nor claims to be true; saying “homo est animal is true” implies a “second intention” which the former sentence does not say anything about.

This solution approaches the complete explanation of all the problems connected to TNT.

2. Buridan versus Augustin

Buridan’s analysis³ could be considered in a way definitive, for an age in which no difference between use and mention, type and token, proposition and sentence was acknowledged. By the way he uses a translated version of TNT which is more judicious, i. e. “no proposition is true” (NPT), so avoiding any ontological complication⁴.

Buridan asserts that we do not need to accept the equivalence $TNT \equiv TTNT$ (or $NPT \equiv TNPT$) in order to refute TNT (NPT). Actually, if “no proposition is true” is true, it is false (there is at least one proposition which is true), while if it is false it is false (there could be true propositions). This is the obvious difference between TNT and the Liar: TNT is self-refuting, while the Liar is paradoxical.

He also noted that if NPT is the only existent proposition it can appear to be true and false at the same time. We may give a set-theory translation: if NPT means “the set T ($t_0 \in T$ iff $t_0 = \text{true proposition}$) is empty”, NPT cannot belong to NPT, so it must be false, and if NPT is false, being NPT the only existent proposition T is empty, so NPT is true. We may call this “epistemological Liar”.

But finally Buridan points out this do not change the result, as the conjunction “NPT & NPT is the only proposition” must be necessarily false.

Probably the most interesting aspect of Buridan’s analysis as to the problem which I would like to point to (i. e. the possible link between the refutation of TNT and some form of ontological factualism) is the conclusion. In all his *Sophismata*, Buridan claims that in logic we are interested only in truth and falseness insofar as they create contradictions, but we are not concerned by the effective existence of God, or whether some necessary truth must be found (p. 242). Anyway, Buridan does not imply any form of formal coherentism, on the contrary he is decidedly in favour of a foundation of logic on reference (pp. 46-47), and this is the typical blend of empirical late medieval nominalism. Actually, he says: “as to what is nothing, there is nor

³ See John Buridan, *Sophisms on Meaning and Truth*, ed. T. K. Scott, Appleton, New York 1966.

⁴ “x does not exist” implies the set of problems created by non-denoting nouns, and it is a form excluded in some logical systems: namely Lesniewski’s Ontology)

being in a way neither being in an opposite way" (p 46), and he elsewhere specifies: "an affirmative sentence whose subject has no reference is false, and this is valid but for *nominal* propositions".

So he wants a referential logic, but not a *metaphysical* one (he does not want – so to say – any ontological commitment). Hence the only possible solution is the idea of possible worlds: but Buridan only touches this solution, without entering into it. He says that NPF is *possible*, i. e. we are able to understand it, and see the state of affairs which could make it true, but "when we formulate it", he says, "it immediately ceases to be true". Here the late medieval procedure of "casus" is involved (p. 46), i. e. the ontological implication of the problem of truth. But the most interesting thing to note is that Buridan's solution, pointing to the "possible" state of affairs which can make TNT be true, clearly evokes a *modal* ontology.

3. Elenchos

Anyway, Augustin's refutation of TNT has a particular structure that is worth noting, as it corresponds to a very common procedure. It is an *elenctical* refutation, analogous to the one Aristotle uses in his counter-refutation of the non contradiction principle. Elenchos is a sort of refutation by means of self-reference: I refute a thesis by showing that it cannot be applied to itself. Karl Otto Apel⁵ and more recently Thomas Nagel have brushed up this old procedure in their criticism of contemporary nihilism, subjectivism, relativism, contextualism and so on.

Is the thesis "there is no objective, universal, univocally valid principle" an universal, objective or universally valid principle? Obviously it cannot be, so such a thesis or other similar ones, are to be considered out of interest. We may note that the elenchos of truth is different from the equivalence thesis mainly because it seems to be intentional while the latter is extensional. Truth is not a simple operator: it is a horizon, a condition, a performative premise...

But also according to these specifications or whatever else, elenchos has some equivocal and problematic aspects. Firstly, it seems to be only "formal", i. e., it shows that some logical-linguistic structure (or construction) has some logical-linguistic consequences and effects, but it does not say anything more. This immediately shows a certain similarity (or parenthood) between the elenchos of truth (ET), that is:

$$\text{TNT, TTNT} \therefore \Box \neg \text{TNT} [\text{or } \Box \text{ T}],$$

⁵ See for instance: K. O. Apel, *Transzendente Semiotik und die Paradigm der "prima philosophia"*, in Benjamins B. V. (ed.), *Integrale Linguistik*, Amsterdam 1979.

and the ontological argument of God's existence (OA) – a similarity which has been immediately noted by St Augustin's argument's commentators. So we may object to it using the same objections typically opposed to OA. Stated that some structure called "truth" is logically undeniable, it does not follow that effectually some truth must exist or must be conquerable in any way (see on this problem Baldwin, 1991). Analogously from the logical necessity if God does not follow his ontological necessity (his existence).

But even as a formal procedure, the elenchos has some evident ambiguities. It is actually a reflexive, self-referential procedure. It implies that (in the case of truth) the sense of "x is true" in "TNT is true" belongs to the same truth which TNT declares to be non-existent. So the exact formulation of TTNT should be:

$T_1 \text{ TNT}_0$ (it is true₁ that there is no truth₀)

This is by the way the most common argument used to defend *nihilism*. Many Nietzsche's commentators hold nihilism to be a combination of TNT + "there are no universally valid principles" (see Schacht, 1996): hence a totally self-refuting theory (TNT is false + contingent). But it implies probably something more, as we see now that the exclusion of the *predicative self reference* of truth (the claim that no truth₀ may be said about truth₀), beyond any expectation, gives some advantages to nihilists' thesis, namely: it gives them the possibility of locating themselves in a field which is somehow beyond the rules of logic. If "x is true" may be plurivocally used, one could admit that the truth of statements like TNT, or "there is no objective truth" or other relativistic, contextualistic and similar stances is different from the truth which they "are about".

4. Mutual overcoming of undeniable entities

Evidently in order to avoid the self-referential interpretation of TNT some paraphrase of TNT is needed. For instance, Rorty (1998) has argued that TNT should be properly intended as "truth is not a goal of inquiry" (TNG), but this could be really misleading, probably more than decidedly admitting TNT. If there is any possibility of getting some sort of truth (if some truth could exist) then it is difficult to agree with Rorty, claiming that TNG.

On the opposite side Nagel (1997) has argued against the possibility of denying truth and objective principles from a point of view not submitted to the rules of reason (so he argued against the possibility of staying out of reason in order to criticise it). He seems to disagree even with the translation of TTNT into some $T_1 \text{ TNT}_0$, i. e. with the idea of a *translogical* logic committed to criticise logic, or a *transrational* point of view committed to criticise reason.

But what is “reason”? What are its bounds and borders? Nagel, quite ingeniously, uses to *define* reason the same arguments commonly used to *defend* it, i. e. the elenctical procedures. Reason is the realm of undeniable thoughts, thoughts that we cannot deny without contradiction (Apel called this type of thoughts “Letzbegründungen”), and these thoughts are: the existence of self, the mathematical truth, the existence of thought. Nagel’s theory has a number of obscure points. Above all: are we sure that these undeniable thoughts (or principles, according to Apel) get on well together? Are we sure that there is no mutual contradiction among them?

For instance, history is surely the undeniable foundation of mathematical truth according to the historicists, as even mathematics, in the form in which we usually practice it, is a historical and human product. So *history is an elenctical concept*: we cannot deny history, as *even our negation will be placed into history*. The same ET holds also for History (and for Being, Self, Language, World: Rosenzweig called them “irreducible” entities). But on the contrary: is there any recognisable history without the *numerability* and individuality of historical facts and events? So *also mathematical truths are undeniable*: as they are *the same transcendental conditions of any understanding of history*.

All that established, what is there (if there is one) the *supreme* undeniable entity? Is mathematics the foundation of history, or is the latter the foundation of the former? Probably the best attitude to adopt is an anti-skeptical perspective, like the one performed by “objective relativists” such as Dilthey or Simmel. They considered truth as having the exact requisite of *being excessive*, not only due to its ubiquity, as the deflationists say, but also due to the “natural” *multiplicity of its possible forms* (historical truth and mathematical truth, for instance, tend to mutually overpass themselves).

But this ultimately means that (as Nagel claims) we cannot properly go out of reason in order to criticise it, because we need the rules of reason even to formulate our criticism, but we cannot properly stay in it, because there are so many contradictory rules inside, that each time we have to decide how to use them (mending the boat while sailing, probably means substitute one undeniable foundation with another, changing the rules....).

5. Elenchos and factualism

History may be the undeniable foundation of thought, but also mathematics may be the undeniable foundation of all possible thoughts. And also: language may be the foundation of thought and reason, and thought also may be the foundation of language, history, reason and so on. Reason in any sense, and being (in Heidegger’s sense) may be the foundations of thought, history and language, but anyway there is no being without history, that is without time, and at last we are thinking of and talking about something called “being” only because we have a language which so to say “creates” the word

“being”.... So there could be: a *historical* reason, a *mathematical* reason, a *cognitive* reason and an *ontological* reason. But finally Reason, in the proper sense, ends to be only this wide and contradictory realm of mutual overcoming of many different forms of reason... So the relativist seems to be right anyway, even if he does not adventure “out” of the realm of Reason.

We may note that this has some relation with our problem, as it depends on the fact that all these elenctical entities are not so firm and strong in their own foundational properties. At closer sight, elenchos does not find an ultimate point (see Galvan 1995): it could be stopped only thanks to a thematical decision. In a simplified manner, and as a last result, if the “truth” which TNT is about is the same truth that we use to refute TNT, we should also use TNT in order to refute TTNT. That is: stated that TNT is a true sentence, it denies the existence of true sentences, so also TTNT is false. Buridan and Duns Scotus agree on this point. TNT has two possible consequences: that *really* no truth exists, so we cannot say anything true or false about whatever subject, or that it is really *true* that no true exists, so at least some truth must exist.

But it is quite obvious that, as we also saw in Buridan’s analysis, a complete elenctical refutation of TNT needs some ontological premise. St. Thomas Aquinas plainly admits that TNT cannot be true, because there is being, there are things, so truth exists. Duns Scotus and Buridan accept it but in a negative way: in order to assert the perfect equivalence of TNT and TTNT we need to suppose a world which necessarily exists, because if we have no being and no thing about which formulate true statements, truth could be completely denied. As we saw, all that according to Buridan means that TTNT is different from TNT and the way towards a definitive refutation of TNT is another one.

So a certain metaphysical sureness about the effective and necessary existence of things is required in order to make use of ET. If there are no being and thing, neither the *intellectual* nor the *factual* possibility of thought or history or language or truth could be admitted; so the elenchos of truth and probably any kind of elenchos are based on some form of factualism. Against the objection of Scotus, Cajetan (1468-1534)⁶ replies that even in the absence of every thing and intellect, a single sentence remains, which is true: the truth of “nothing is nothing” fundamentally *remains*, at least because *if there were any intellect*, it would be compelled to follow that rule.

This is the point of the problem, I suggest, as the obvious objection to Cajetan would be, anyway: who ever could pronounce or think the “rule”, i. e. the proposition “nothing is nothing”? As we will see, an important ontological decision is involved here.

⁶ In *Summam Theologiam*, I, 2, ad. 3, n. 11

So factualism supports elenchos. But correlatively, elenchos supports factualism. Actually, how can we assert the necessary existence of being without supposing that even the negation of being happens-takes place *within* being, i. e. without supposing being as a all- surrounding horizon, which contains everything (as elenchos says and requires)? Elenchos is the strongest and most typical meta-ontological procedure: the beginning and the foundation of Aristotle's *Metaphysics* itself, asserting the *ubiquity* (and the multivalence) of being, is the anticipation and foundation of elenchos.

6. There might be nothing

The elenctical refutation of TNT requires some form of factualism, and more precisely the negation of the eventuality expressed by "there might be nothing" (TMN): $\neg$ TNT requires $\neg$ TMN.

Correlatively, we may also admit that whatever acceptance of TNT needs to be supported by the preliminary acceptance of TMN. But is TMN a thinkable hypothesis?

In terms of a theory of possible worlds (see Van Inwagen: 1996, Lowe: 1996, Baldwin: 1996) TMN could be interpreted as the possible existence of a "nil-world", a world in which there is nothing. Such a world seems to be intimately contradictory, as "world" is $=_{df}$ "totality of things", and there might not be such a totality without things (see Lowe: 1996). Besides, according to van Inwagen, the possible existence of a nil-world, even admitting its so to say "logical", "analytical" possibility (for instance changing the definition of "world"), is fundamentally null, like any other ontological possibility. If the possible worlds are infinite, each of them has 0/infinite probabilities of being existent, that means 0 probabilities.

Baldwin has challenged this theory, substantially showing that the premise, i. e. van Inwagen's equi-probabilism, is wrong: there are some worlds that, with respect to the real world, are "more probable" than others. For instance: the world in which Moore dies at the venerable age of 145 is less probable than the world in which it dies at 70 or 89. Baldwin also finds a way to demonstrate the ontological possibility of the nil-world, based on a "subtraction" argument. The nil-world is obtained from the real world by taking away each thing or element whose non-existence does not imply the existence of some other thing or element (see also Baldwin's method "implementation", by). When all such things are taken away, we will have got something like a min-world, in which there are only ontologically connected things. Hence, taking away only one thing from this min-world we will obtain an effective nil-world: or rather the world whose absence does not imply the existence of any other thing or people.

This conclusion could sound odd, as Baldwin effectually constructed some sort of absolute "nothing", i. e. a nothing which is ontologically independent of any other thing or people, whose non-existence does not imply

the existence of any other thing. That means, in a way: a construction of a “non-thing” which does not need to be constructed. This circularity, or *hysteron proteron*, is a typical feature of rationalism (Descartes in his ontological proof of God’s existence and in his deduction of Cogito used a structure which is somehow similar to this). But we may note that in this case all the process is developed in a negative way. “The” nothing anticipates our search and/or construction of it, even if we may obtain it (we may see it) only by means of a subtractive ontology. We can call this a “negative elenchos”. With the positive elenchos, as we saw, one obtains an ontological background which is intended as the factual, real world of things that are around us; with the negative elenchos we deal with something which is here (is constructed here) but refers to another world: it refers to a somewhat “non-being” which does not need to be destroyed or denied in order to be non-existent.

7. Three scenarios

So now we are able to come to our conclusion. Three ontological scenarios are opened according to our last considerations. The first is the one envisaged by van Inwagen’s account (or some other form of absolute pluralism): the infinite series of possible worlds in which the real world and the nil-world are only two of the possible worlds. Schematically:

$w_s \ w_t \ w_q \ (W_r) \ w_1 \ w_2 \ w_3 \dots W_{nil}$

W_r is the effective real world, and $w_s \ w_t \ w_q \ w_1 \ w_2 \ w_3 \dots$ are the other possible worlds, while W_{nil} is the nil-world. There is no reason why any possible world should be preferred, not even the real world as such shows any reason for why it should be in this way or in another. Besides, there might be infinite criteria for the probability of worlds, as each world could be considered as paradigmatic.

A factualist can easily contrast that picture by asserting that anyway all these worlds are thought and imagined and contained in this real world in which we are. So the factualist admits a scenario like this:

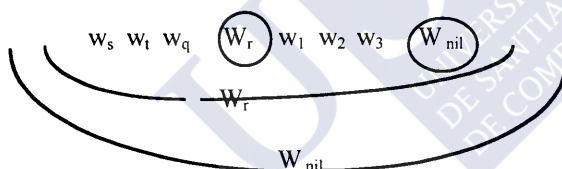
$\underbrace{w_s \ w_t \ w_q \ (W_r) \ w_1 \ w_2 \ w_3 \dots W_{nil}}_{W_r}$

All possible worlds are contained in the real world. The last also offers them some criteria of etero-probability. We may also note that the real world is double: it is the real world in which each possible world is included, and it is one of the possible worlds. But this way two forms of factualism are possible. The first is: the real world is alleged to be double, as it contains in a way its own image; the second is: the real world does not properly contain its own

image, but only the image of possible worlds (it cannot actually include its image without infinite recurrence). But evidently the second type of factualism (which in a way could be compared to Kant's refusal of the "total" world image in KrV) is contradictory, because among the possible worlds there might be some world which differs from the real one for only one point: because it includes itself. So the self-referential world should be excluded not only on the meta-theoretical level of the scheme, but also in the modal theoretical level of the analysis of the possibilities. This is properly a form of rigorous factualism, but we see that it is based on a preliminary restriction: a consequent factualist should prevent her or himself to think of some contradictory ontological possibilities.

As we see, among the possible worlds the nil-world is evidently included. But it is fairly consequent to admit that the nil-world also stays around the world in which we are, as we do not know if the effective universe and world are infinite or without a limit, beyond which there would be "nothing" in the proper sense. So there must be a nil-world, a world in which there is no thing and being, even around the real world that contains all possible worlds.

Hence the third and last scenario is as follows:



There is a duplication of the nil-world, as in the case of the real world. The world-nothing is one of the possible images of the world, but it also stays in the background of all possible image, it also "contains" the real world, at least because it surrounds it. So there is a similarity between W_{nil} and W_r : they are both back-images: insofar as images, they stay inside and outside themselves, they stay in front of the ontological perspective, in front of the sight that considers the ontological possibilities, and at the same time they are condition of any sort of sight, they are (they can be seen as) surrounding and founding any transcendental condition, any possible ontology. We should remember that when we say "real world" (being) and "nil world" (nothing) we mean these complex configurations which are inside and outside itself

Now the question is: is one of these three scenarios to be preferred? If we choose the first, we will have some sort of pluralism and relativism, if we choose the second, we will get some form (mainly: two possible forms) of factualism; if we choose the last scenarios, we will imply some sort of nihilism as an ontological foundation of logic, knowledge and truth (TNT + TMN). But what about a general consideration of all the three scenarios?

What sort of ontological trend is needed to catch them and see their possible developments?

We may admit that the three possibilities belong to the general field of *modal ontology*. So I will call the fourth – referring to the medieval concept of “transcensus” – a sort of “trans-ontology”, or a sort of “trans-platonism”, as the three possibilities obviously belong to the realm of what Plato and Platonist tradition call “pure thoughts” (and by the way Hegel called “das Logische”, intending the pure realm of thoughts). This perspective is historically and linguistically determined, as the emergence of the three possibilities can be noticed *in the history of philosophy and in the general use of philosophical language*. This casts, I suggest, a new light on the development and reasons of modal ontology, and in a way it may support it, against the objections of anti-modalists and factualists.

For instance, in his criticism of modal ontology Nicholas Rescher (see: 1999) observes that possible worlds are “uncountable”, not because there are so many of them, but “because they cannot be identified”; “we cannot count what we cannot individuate” (p. 420). But this also means that there is no possible “obstensive confrontation” of such possibilities, hence no form of *de re* possibilism is possible. Rescher here introduces the typical formula of immanentism: “we have no way to get there from here”, we cannot “get out”, even fictionally, from our world. “we can indeed operate a possibilism that is *de dicto*”, but “what we then can and do have is conceptualism: an ‘ontology’, if you will, of conceptualizable possibilities” (p. 419).

It is not so clear in Rescher’s discourse what could be a modal *de re* ontology, and it is not so clear in which sense his criticism really hits modal fictionalism. He asserts that “possible worlds should be dismissed from the ontological sphere” (p. 414), but this way he clearly shows he agrees with a Quinean but very arguable identification between *ontology* and *factual ontology*: he seems to consider the study of what there is and what it is like perfectly cohestensive to the study of being in general. According to a loosely Meinongian account it is fairly reasonable to call factual ontology “metaphysics”, but even in Lowe’s or Jackson’s point of view (see Lowe: 1998; Jackson: 1998), where “metaphysics” is used in a non-European sense, the study of “what there is” is only a part of “first-order metaphysics”. Obviously, in modal analysis we are dealing with “thought-objects”, and the question is: what is the “ontology” of this thought objects? Are they possible objects of some meta-ontological analysis (see. Van Inwagen: 1999)? And this is, I suggest, properly the problem of the late-Platonism (from Hegel to Natorp and Gadamer). In a way, the trans-Platonism that we indicated is a part of a so-to-say “third-order metaphysics”, as the three ontological possibilities that we found are general philosophical conditions, but in another way, they are also simple structures or conditions of looking at and talking of things, so they should belong to a “second-order metaphysics (or ontology)”. Finally, in

a third sense, they are objects of historical and meta-philosophical reflection. So they also belong to an ontology of historical and philosophical "objects".

All that may also suggest some indicative conclusions about truth. We found that truth is fundamentally reflexive, "more than logic", and in a way more than "true". Now we may add that truth is a "trans-concept": a concept which is located on the borders of the meta-theoretical and "encyclopaedic" distinctions, because in a way is the same condition of those distinctions. The possibility of catching all these aspects of truth is linked to the possibility of the trans-platonism above described: a perspective which also tends to cross and overpass the meta-theoretical distinctions, locating itself on the level of historical development of philosophy.

REFERENCES

- Thomas Baldwin: "There Might Be Nothing", *Analysis*, 56, 4, Oct. 1996
- W. Charron, J. P. Dole: "On the Self-Refuting Statement 'There is no Truth': A Medieval Treatment", *Vivarium*, 1996
- Harty Field: "Epistemological Nonfactualism and the A Prioricity of Logic", *Philosophical Studies*, 92, 1-2, Oct. 1998
- Sergio Galvan: "A Formalisation of Elenctic Argumentation", *Erkenntnis*, 43, 1995
- Sergio Galvan: *Aspetti formali dell'argomentazione filosofica: analisi del regresso all'infinito*, in F. Castellani e L. Montecucco, eds., *Normatività logica e ragionamento di senso comune*, il Mulino, Bologna 1998.
- Peter van Inwagen: "Why is there anything at all?", *Proceedings of the Aristotelian Society*, Supp. 70, 1996
- Peter van Inwagen: "Modal Epistemology", *Philosophical Studies*, 92, 1-2, 1998
- Dale Jacquette: *Meinongian Logic. The Semantics of Existence and Nonexistence*, de Gruyter, Berlin-New York 1996
- Frank Jackson: *From Metaphysics To Ethics*, Clarendon, Oxford 1998
- E. J. Lowe: "Why is there anything at all?", *Proceedings of the Aristotelian Society*, Supp. 70, 1996
- E. J. Lowe: *The Possibility of Metaphysics*, Clarendon: Oxford 1998
- Thomas Nagel, *The Last Word*, Oxford University Press, Oxford New York 1997
- N. Rescher, "How Many Possible Worlds Are There?", *Philosophy and Phenomenological Research*, LIX, 2, June 1999
- R. Rorty, *Truth and Progress, Philosophical Papers*, vol 3, Cambridge University Press, Cambridge 1998.

Franca d'Agostini
 University of Turin
 e-mail: filfda@tin.it

PHYSICALISM AND THE FALLACY OF COMPOSITION

CRAWFORD L. ELDER
University of Connecticut

ABSTRACT: Physicalism, as in this paper, holds that every instance of causation reported by the special sciences is shadowed, even rivaled, by causation at the level of microphysics. The reported “cause” is embodied in one massive collection of microparticle events; the “effect” in another; the former brings about the latter by laws of physics. This paper argues that while individual events in the “cause” collection bring about individual events in the “effect” collection, it does not follow, and is false, that the “cause” collection itself brings anything about. Causings reported by the special sciences can be traced only at that level.

KEYWORDS: Physicalism, causation, reductionism.

A mutation alters the hemoglobin in some members of a species of antelope, and as a result the members fare better at high altitudes than their conspecifics do; so high-altitude foraging areas become open to them that are closed to their conspecifics; they thrive, reproduce at a greater rate, and the gene for altered hemoglobin spreads further through the gene pool of the species. That sounds like a classic example (owed to Karen Neander, 1995) of a causal chain traced by evolutionary biology. But a view now nearly universal among philosophers maintains that such biological causation is always shadowed, perhaps even rivaled, by causation on a different level¹. That the subgroup of antelopes forages in areas closed to the conspecifics is a state of affairs *embodied* or *realized*, notes this view, in certain movements and state changes done by certain physical microparticles—untold billions of microparticles and movements, but a finite and determinate (more on this below) collection nevertheless. That the subgroup reproduces at a greater rate is likewise realized by a huge collection of microparticle movements, a

¹ The view most often is articulated as a claim that *mental* causation is shadowed, even rivaled, by causation at the level of microphysics. But arguments for that version of the view commonly entail that not just mental causation, but causation reported by *any* of the special sciences, is thus shadowed. See Sturgeon 1998, e.g. at p. 418 (“a physicalism that is both general and severe”), and cf. Yablo 1992a.

different collection. And the microparticle happenings comprised in the first collection are causally responsible, *strictly in accordance with the laws of microphysics*, for the microparticle happenings in the second. Biological causation is always shadowed, perhaps even rivaled, by causation on the level of microphysics.

The view I mean is general: any case of causing uncovered by any of the special sciences can be recaptured at the level of microphysics. This view is I think what most philosophers mean by “physicalism”; in any case, “physicalism” is the label I shall use. Physicalism comes in two forms. Modest physicalism holds that any causal transaction reported by the special sciences can be retraced by microphysics². Hegemonic physicalism holds that retracing such a transaction at the level of microphysics is actually reporting the *only* causation that is going on: the happenings identified by the special sciences only appear to cause something, and the real causing is done by large collections of microparticles³. This paper focuses on the weaker thesis of modest physicalism. But I will bear in mind that it is perhaps less stable⁴, and certainly less exciting, than hegemonic physicalism.

This paper argues that modest physicalism lacks supporting argument; a fortiori, hegemonic physicalism lacks it as well. There is no reason to suppose that causings reported by the special sciences are in general shadowed by causal transactions between one collection of microphysical events and another. This paper attributes the impression that physicalism is undeniable —modest physicalism, at least— to the fallacy of composition. It is undeniable, let us allow, that each microphysical happening comprised in a “cause” collection causes *some* further microphysical happening, and we can temporarily grant (though see below) that each such microphysical happening

² D. M. Armstrong espouses a modest physicalism (see, for example, 1997, p. 153), though some of his formulations may sound like hegemonic physicalism instead (e.g. 1997, p. 6 or p. 253).

³ Jaegwon Kim’s arguments are among the most forceful there are for hegemonic physicalism, but Kim’s own position is somewhat hedged: either the causings reported by the special sciences really are done by collections of microparticles instead, or there is no question of “instead” —the causes identified by the special sciences *are* just a matter of such collections of microparticles coming to satisfy very complex microphysical descriptions. See Kim 1989a and 1989b, or Kim forthcoming.

⁴ Less stable mainly because of the problem of coinciding objects. If the medium-sized objects posited by the special sciences actually *cause* things, and figure in laws, they have as valid a claim to real existence, *prima facie*, as do the microparticles which compose them. But the parcel of microparticles which (it is said) composes any such object has very different modal properties from that object itself —medium-sized objects are generally “mereologically incontinent”, while parcels are “mereologically stable” (see, e.g., Zimmerman 1995). Modest physicalism may be less stable also because of the problem of harmonizing laws. If individual microparticles are bound only to obey the laws of microphysics, how can it be that parcels of microparticles cannot fail to obey laws belonging to different sciences altogether?

causes, by virtue of a causal chain which it launches, some individual microphysical happening comprised in the corresponding "effect" collection. It does not follow that the cause collection causes all the happenings in the corresponding effect collection, or causes a large subset of them, or causes anything. Separate argument is needed to establish those claims. None is available, I shall suggest.

This paper trades on three assumptions about causation which there is not room here to defend—but they are fairly common assumptions, and physicalists especially should be friendly towards them. The first is that the relata of causation are finely-individuated entities. I have argued elsewhere that they are states of affairs (Elder forthcoming). For the purposes of this paper, it would be just as good if they were finely-individuated events, of the sort which Jaegwon Kim (1969, 1980) discusses or the sort Stephen Yablo (1992b) defends. What they cannot be, for the purposes of this paper, are coarsely individuated events such as Davidson (1967, 1969) affirms. The second assumption is that every individual causing instantiates some general causal law. The third is that for a state of affairs (or Kim-event) *c* to be a cause of effect *e* is for it to be what Bennett (following Mackie) calls "an NS condition" of *e* (Bennett 1988, Ch. III). That is, *c* must be, in a way set by the causal laws of nature, a Necessary (i.e. indispensable) component of developments Sufficient, given the surrounding circumstances, for occurrence of *e*.

Physicalists often elect the hegemonic variant of their position, and to this extent they must be favorably disposed towards the first two of these assumptions. Without the first assumption, the collection of microparticle happenings which realizes the wider foraging by antelopes in the subgroup, and that wider foraging itself, are really one and the same event, just differently described. So whatever the former causes, the latter causes. Distinguishing these events (or states of affairs) is what opens the door to the claim of hegemony. What substantiates that claim, for many physicalists, is the further thought that the laws which underwrite causation at the level of the microparticles are far better, as laws, than the laws (*putative* laws, perhaps) which underwrite causation by the happenings identified by the special sciences. The laws of the special sciences, hegemonic physicalists argue, are riddled with exceptions unless protected by *ceteris paribus* clauses. Yet the *ceteris paribus* clauses seem in principle to resist precise specification, unless as "except where this laws fails to hold". This suggests, however, that the individual causings reported by the special sciences are less than genuine only if causings *need* to be underwritten by causal laws. So physicalists tend, and should tend, to be well disposed towards my second assumption as well.

I

But my conclusion may seem too sweeping, and not just to physicalists. Does it really never happen that an instance of causation at the level of medium-sized objects is underlain by an instance at the level of microphysics? In one particular sort of case, I admit, this does indeed happen. The burden of this paper is just to establish that it does not always happen—not in all individual causings reported by what physicalists call “the special sciences”.

The sort of case in which it happens is the case in which there is a rationale, *from the standpoint of microphysics*, for collecting together all the microparticles which figure in realizing the event identified as “cause” at the level of medium-sized objects. There rarely is such a rationale, I shall argue, with the individual “cause” events reported by biology or economics (or, for that matter, sociology or behavioral psychology, etc.)

To say what such a rationale would look like, and why it matters, will require sections II and III of this paper. But a picture here may be worth many words later. I nudge an irregularly shaped rock towards the edge of a table. By pushing its center of gravity beyond the edge, I cause the rock to fall. So, at least, mechanics—and what some call “folk physics”—reports. Can this instance of causation be recaptured at the level of microphysics? That the rock’s center of gravity is at t_1 above nothing but air is a state of affairs realized in the positions and motions of a myriad of microparticles. The rock’s fall at t_2 is embodied in a different array of positions and motions, involving very largely the same myriad of microparticles. But plainly it does make sense, from the standpoint of microphysics, to take all the microparticles that figure in realizing the “cause” event *together*. For the fate of each is intertwined with the position and state of all the rest. Wherever the avalanche of microparticle events begins, that terminates in the “effect” event—in the careers of whichever microparticles—that avalanche ends up embroiling all other microparticles that figure in the “cause” event.

But with typical instances of causation reported by biology, say, or by economics, this synergy—this acting in concert—disappears at the level of the microparticles. Or so I shall argue. Consider, for example, the high-grazing antelopes. First we need to determine which microparticles it is, whose movements and state changes jointly embody the “cause” event—the state of affairs that these antelopes now forage in places not open to their conspecifics. This question indeed raises puzzles of its own. The start of the answer seems to be: all those microparticles contained in the bodily boundaries of the high-grazing antelopes and their conspecifics. For otherwise the physicalist would run a risk of not circumscribing enough microparticle happenings to capture the fact that it is *antelopes* which differ in how they are faring. But there will then be familiar problems about just which microparticles to count as falling within these boundaries, and which without

(Unger 1980; cf. van Inwagen 1990, p. 214 ff). How, in terms specific enough to generate a list of microparticles, is it determined where a hair in the antelope's coat leaves off and a drop of water that has begun to permeate that hair takes up? Perhaps the question can be finessed by generosity. Let the physicalist include all molecules that lie within a one-inch zone outside what common sense regards as the antelope's outer surface; let him also include all microparticles in molecules of gases in each antelope's lungs; etc. But next note that the physicalist must circle enough microparticle occurrences to embody the fact that the genetically well-favored subgroup is *foraging* unusually well. Successful foraging involves the ingestion of nutrients. So the microparticles in all the nutrients, contained in all the high-meadow grasses that get ingested, must be included. Then too, what about the jaw movements by which these antelopes tear and chew these nutritious grasses? If the physicalist is to capture the fact that these are applications of the antelopes' normal behavioral repertoire in what is merely a new setting, perhaps enough of the evolutionary history has to be included to capture the fact that the antelopes have foraging in their behavioral repertoire—that neural routines for foraging were at one time selected for (Millikan 1993). Now the cause collection of microparticle movements starts to stretch back in time to ancestor antelopes. How likely is it that there will be a rationale, *from the standpoint of microphysics*, for collecting together the states of all these scattered microparticles as a "cause" event?

If the typical causal transactions reported by evolutionary biology look messy and sprawling, when viewed at the level of the microparticles, so too, famously, do the causal transactions reported by economics. Consider any case in which economics depicts, as the cause of some individual outcome, the fact that money is distributed in a particular economy in a particular way. Money is always physically embodied. But, as Jerry Fodor points out in a different context (Fodor 1975, p. 15), there seems to be virtually no limit to the sort of physical stuff of which money can be composed. Its status as money depends not on its physical composition but on something relational and social—on its being sanctioned to play certain roles by a society or a government. So again there is a hard question about how even to *begin* assessing the physicalist's claim—how to determine just which microparticle events fall within the shadow of the monetary distribution which economics depicts as a cause. The physicalist must assemble enough microparticle movements to recapture, at the level of the microparticles, the fact that *money* in this economy is distributed in such-and-such a way. For this it would evidently not be enough to count just movements of all microparticles *in* all actual pieces of money. One needs enough microparticle events to capture the fact *that* those pieces have an assigned social role. So one has to include movements and state changes in at least some microparticles in some persons, perhaps reaching well back in history. Perhaps also one must include

microparticle events which jointly realize the existence of certain legal documents, certain courts of law, certain offices. How likely is it that there will be a rationale from the standpoint of microphysics for collecting all these happenings together as a cause?

To sum up: true enough, any individual occurrence warrantably depicted by a special science as causing something has a physical embodiment. But the collection of microparticle events which jointly embodies it will typically be sprawling and disjointed, to a striking degree. That point will form the nerve of my argument against physicalism. And already we have one result: it is not really plausible to say that *every* last microparticle event comprised in the physicalist's "cause collection" is causally involved in bringing about the collection of microparticle events which shadows the "effect" explained by the special science. Some of them are just along for the ride, it seems. (More on this in section III.) Still, this result need not damage the physicalist's central claim. We must still ask: if the physicalist's "cause collection" is often bigger than it needs to be, to cause what it is supposed to cause, is it always at least big enough? Is it causally sufficient to bring about all the microparticle happenings which jointly realize the outcome which the special science claims causally to have explained? If so, the heart of the physicalist's position is intact.

II

Should we think of the physicalist's "cause collection" as producing the corresponding "effect collection" all at once, or component microparticle event by component microparticle event? In the present section I shall consider the former suggestion, and in the next section the latter.

Imagine then that the microparticle events which embody the lucky antelopes' foraging together cause, in a single (if protracted) instance of causation, the collection of microparticle events which embodies the fact that the lucky hemoglobin gene spreads thorough the species. But how can this be, if every individual causing must instance some general law of nature? Could there really be a law of nature to the effect that whenever such-and-such movements and state changes befall microparticles of such-and-such a description—the list of these being billions of places in length—other movements and state changes, again in the billions, will befall billions of other microparticles?

The problem is not that such a law would be "general" in name only. It *would* be general in name only, of course. Any law which subsumed the antelopes' high-altitude foraging under a description as fine-grained as this would assuredly come into play only once in the history of the world. (It would be "exceptionless" in only an empty sense.) Nor is the problem that, because antecedent and consequent in such a law would get satisfied only once in history, we could never have warrant for believing it to be a law. For

plainly the idea would be that we were warranted in a different way in believing this. The idea would be that the massively specific law is a *theorem* jointly entailed by simple and familiar laws of microparticle causation —laws which, for example, link a movement or state change in one microparticle with movements and state changes in a few other microparticles. Take several such simple laws together, the thinking would run, and plug in a massively complex description of antecedent conditions (capturing all or most of the “cause collection”), and out comes a massively complex description of an effect.

But why think that in addition to the simple and truly general laws of particle physics which jointly entail that theorem, that theorem is a law of nature in its own right? Such a position seems to run counter to actual scientific practice. Suppose that a given science affirms a number of separate causal laws governing phenomena within its domain. The separateness of the laws then depicts the transactions covered by the laws as causally independent of one another. Will that science also take there to be a law governing what happens when the antecedent in one of these separate laws takes one value, *and* the antecedent in another takes a second value, *and* the antecedent in a third takes a third value? Sometimes, but not always, and not automatically. If it turns out that specific values of the independent antecedents sum together in distinctive and characteristic ways —as when gravity and electromagnetism combine to shape the trajectory of an ion on the surface of the sun— then the science will judge there to be a separate law governing conjunctions of values for antecedents in each of the original laws. If it turns out that values of what looked like quite different antecedents are really variants on a single underlying phenomenon —think of the reductions of gas laws to molecular kinematics— the science will posit a new law or laws, and will take the original separate laws to have been theoretically reduced. Either way it is crucial that the new law predict and explain phenomena not already predicted by the conjunction of the original separate laws alone.

Now we have so far taken the physicalist to hold that the microparticle events composing the effect collection get produced by causal chains largely independent of one another, starting from causally independent microparticle events in the cause collection. (More on this in the next section.) Each of the causing microparticle events shoots its causal influence down to one of the microparticle events in the effect collection, and the shots do not (in general) cross. No large-scale summing of causal sources occurs. No massively complex law of microphysics is needed to predict all the events in the effect collection. If such a law were formulated, it would predict nothing not already predicted by the laws which jointly ground it.

I conclude that the physicalist has no justification for regarding the massively complex generalization, which would link the cause collection as a whole to the effect collection as a whole, as a law of nature. But I have

assumed that every individual causing must be underwritten by a law of nature. It follows that what we have called “the cause collection” does not after all cause “the effect collection” —not if we think of this causing as a one-shot causation of “the effect collection” as a whole.

III

But the physicalist can, I have just allowed, maintain that individual microparticle movements comprised in “the cause collection” bring about, via separate causal chains, each of the different microphysical happenings that compose “the effect collection”. One might ask: isn’t this enough? If the physicalist depicts the causing done by “the cause collection” *as a piecemeal affair*, involving many separate instances of causation, could she not rightly claim that in effect the cause collection produces the whole effect collection after all?

The problem with this suggestion lies not with the laws that would underwrite the causation. The laws invoked would be simple, standard laws of microphysics. For separate microparticle events within the cause collection launch, in accordance with such laws, separate causal chains, each terminating in a separate microparticle event within the effect collection. Between them, I have allowed, the microparticle events in the cause collection are responsible for every event within the effect collection.

No, the problem would lie in claiming, on this basis, that the cause collection itself —*as a whole*— causes anything. It does not help to say that the cause collection exercises piecemeal causation, causing by one chain Microparticle Event #1 within the effect collection, by another chain Microparticle Event #2, by yet a third chain #3, etc. The problem is that saying that the cause collection as a whole causes Event #1, just on the strength of a premise that this or that component within it causes Event #1, commits the fallacy of composition. And not only is the claim logically unwarranted; it is implausible. The cause collection in the case of the antelopes, for example, comprises movements and state changes done by microparticles scattered through high-meadow grasses and low-meadow grasses, through the bodies of high-grazing antelopes and low-grazing ones, even perhaps through the bodies of ancestor antelopes long since departed. Why suppose that a collection of states of affairs involving all these far-flung microparticles *as a whole* produces *any* outcome? Why isn’t such a collection too disjointed to act at all?

The challenge is particularly acute if one assumes, as this paper does, that a cause of outcome *e* is an NS condition of *e* —a state of affairs (or event) which is, in a way fixed by the causal laws of nature, an indispensable component of circumstances which jointly suffice for *e*’s occurrence. For consider any event within the effect collection in our example —any individual microparticle happening *e*, among those which jointly embody the

eventual prevalence of antelopes with altered hemoglobin. To make matters concrete, let e be a microparticle movement in the eyelid of a newborn offspring of a prolific high-foraging antelope. We assume with the physicalist that some microparticle happening(s) in the cause collection launched a many-staged causal chain which resulted in e . But could it possibly be claimed not just that *some* microparticle movements within the cause collection were involved indispensably in launching this one chain, but that *every last one* of them was? Only so can the cause collection *as a whole* qualify as an NS condition for e .

To see just how far-fetched a claim this would be, consider a scenario on which it would after all be true. A causal chain begins just as one of the first high-grazing antelopes bites down on a surprisingly succulent blade of high-meadow grass. What launches the chain is a movement of a microparticle within the antelope's mouth. Successive happenings in this chain involve every last microparticle in every last antelope of that species alive at the time, and in every last blade of grass ingested over the next ten years by any of those antelopes. The chain catches each of these microparticles in just the right state—a state it is in, at the moment the causal chain catches up with it, precisely because of the state it had been in at the moment the chain started. The chain ends up sufficing for the occurrence of event e in the eyelid of the newborn antelope. But indispensable for its causal sufficiency is the state *each* of the scattered microparticles was in, at the instant the chain started. That collection of microparticle states, as complex as it was, and as widely-scattered as the featured microparticles were, was *as a whole* a Necessary component of what Sufficed for e .

This scenario is closer to laughable than to believable. At the macro-level, it seems obvious that the causal chains which lead to the existence of the newborn antelope, eyelid and all, run through the lives of that newborn's forebears; just possibly through the lives, as well, of other antelopes who competed with the forebears for forage or mates; just possibly through the careers, as well, of all the blades of grass which the forebears ate. How astonishing if, on the level of microphysics, the causal chain that leads to e runs far more widely: through the bodies of *every* member of the newborn's species, even completely unrelated ones, and through *every* microparticle in each of these bodies! It would be as if the microphysical chain were exquisitely sensitive to the membership boundaries of the newborn's species, and to the spatial boundaries of those members' bodies. It is far, far more believable that if the microparticle movement in the lucky antelope's mouth does launch a causal chain which results in e , that chain does not embroil every last microparticle featured in the cause collection. No state of that sprawling group of microparticles as a whole—no collection of states each involving one microparticle within that group—is causally indispensable for the production of e . The cause collection is not an NS condition of any

component of the effect collection. So *it* does not cause the effect collection in piecemeal fashion.

IV

This paper argues against a thesis widely regarded as undeniable. Worse yet, physicalism is widely regarded as crucial for our understanding of how the various sciences relate to one another. It would be fair to ask whether an alternative understanding is in the offing, or merely a lack of understanding.

In crude terms —the best I can offer at this point— my alternative understanding runs this way. The objects of the world are composed of microparticles. Microparticles have their ways, but their ways do not exhaust the ways of the objects which they compose. Those objects are genuine wholes, not mere parcels or heaps of microparticles. Their trajectories and behaviors cannot be plotted by summing over the microparticles within them. For they can and do survive defections by individual component microparticles, and accretions of new ones, as mere parcels of microparticles cannot⁵. They have natures of their own⁶.

REFERENCES

- Armstrong, David 1997: *A World of States of Affairs*. Cambridge: Cambridge Univ. Press.
 Bennett, Jonathan 1988: *Events and Their Names*. Indianapolis: Hackett.
 Davidson, Donald 1967: "Causal Relations". *Journal of Philosophy*, 64, pp. 691-703.
 ——— 1989: "The Individuation of Events". In Rescher 1969.
 Elder, Crawford 1998: "Essential Properties and Coinciding Objects". *Philosophy and Phenomenological Research*, 58, pp. 317-31.
 ——— forthcoming: "Mental Causation versus Physical Causation: No Contest". *Philosophy and Phenomenological Research*.
 Fodor, Jerry 1975: *The Language of Thought*. New York: Thomas Y. Crowell.
 Kim, Jaegwon 1969: "Events and their Descriptions". In Rescher 1969.
 ——— 1980: "Events as Property Exemplifications". In Brand, Miles, and Walton, D., *Action Theory*. Dordrecht: D. Reidel.
 ——— 1989a: "The Myth of Nonreductive Materialism". *Proceedings and Addresses of the American Philosophical Association*, 63, pp. 31-47.
 ——— 1989b: "Mechanism, Purpose, and Explanatory Exclusion". *Philosophical Perspectives*, 3, pp. 77-108.

⁵ This may seem to raise what is called the problem of coinciding objects. If objects recognized by special sciences typically are "mereologically incontinent", while the parcel of microparticles which composes any such object is "mereologically stable", we seem to have two objects with different modal capacities occupying exactly the same spot. My solution is that there are microparticles, and medium-sized objects composed of microparticles, but no parcels of microparticles (Elder 1998).

⁶ For comments on earlier drafts I am grateful to Anne Hiskes, Ruth Millikan, Austen Clark, and John Troyer.

- _____. forthcoming: "Making Sense of Emergence". Forthcoming proceedings of the 1997 Oberlin Conference.
- Millikan, Ruth 1993: "What is Behavior? A Philosophical Essay on Ethology and Individualism in Biology". In her *White Queen Psychology and Other Essays for Alice*. Cambridge, Mass.: M.I.T. Press.
- Neander, Karen 1991: "Misrepresenting and Malfunctioning". *Philosophical Studies*, 79, pp. 109-41.
- Rescher, Nicholas et al 1969: *Essays in Honor of Carl Hempel*. Dordrecht: D. Reidel.
- Sturgeon, Scott 1998: "Physicalism and Overdetermination". *Mind*, 107, pp. 411-32.
- Unger, Peter 1980: "The Problem of the Many". In French, Peter A., Uehling, Theodore E., and Wettstein, Howard K., *Midwest Studies in Philosophy, V: Studies in Epistemology*. Minneapolis: Univ. of Minnesota Press.
- van Inwagen, Peter 1990: *Material Beings*. Ithaca: Cornell Univ. Press.
- Yablo, Stephen 1992a: "Mental Causation". *Philosophical Review*, 101, pp. 245-80.
- _____. 1992b: "Cause and Essence". *Synthese*, 93, pp. 403-49.
- Zimmerman, Dean 1995: "Theories of Masses and Problems of Constitution". *Philosophical Review*, 104, pp. 53-110.

Crawford L. Elder
University of Connecticut
e-mail: crawford.elder@uconn.edu



LAS FRONTERAS DE LO FÍSICO*

MANUEL LIZ

Universidad de La Laguna

ABSTRACT: The aim of this paper is to analyze how could it be defined the physical domain. I begin establishing five conditions to obtain an adequate definition. Then, I examine several proposals unable to fulfil those conditions. Here, special attention will be paid to the, so called, layered image of reality. Finally, I propose a new way to deal with the construction of the physical domain. It is based on ostensive references to physical items and on the principle of causal closure of physical properties.

KEYWORDS: Physicalism, physical properties, closure principles, the layered image of reality, second order properties, ostensive definitions, the causal closure of the physical.

1. El fisicalismo: entre la verdad vacua y la falsedad literal

El fisicalismo siempre se ha visto obligado a sortear dos peligros: el de ser vacuamente verdadero y el de ser literalmente falso. Afirmar que todo lo que hay es físico se convierte en una verdad vacía de contenido si se entiende por "físico" algo tal liberal como para acabar pudiendo ser aplicado sin demasiados problemas a todo lo que hay. Y se convierte en una falsedad literal si se entiende por "físico" algo excesivamente restrictivo, algo que en último término no puede aplicarse a nada.

Este destino lo comparten también otras muchas doctrinas filosóficas. Algunas muy cercanas al fisicalismo (como el materialismo o el naturalismo) y otras más alejadas de él (como, por ejemplo, el idealismo). Y es frecuente hoy día recurrir a este rasgo para desprestigiar cualquier posición metafísica con pretensiones de máxima generalidad (véase, por ejemplo, Moulines, 1982: 346-64, así como la polémica recogida en Esquivel (comp.), 1982).

Pero centrémonos en el fisicalismo. La cuestión que quiero discutir es la siguiente: ¿Hay alguna forma de dibujar las fronteras de lo físico de manera que no resulte ni vacuamente verdadero ni literalmente falso pretender que no hay más realidad que la realidad física?

* Este trabajo ha sido financiado con los proyectos PI1998/047 y PI1999/092, de la Consejería de Educación del Gobierno Autónomo de Canarias.

He utilizado a propósito el término "fronteras" porque creo que bajo esa imagen la pregunta anterior sí puede tener una respuesta afirmativa. Y que tal respuesta conduce a una posición metafísica de máxima generalidad que sí puede enriquecer nuestra comprensión de la realidad.

Al igual que ocurre respecto a los cambios de frontera de cualquier país que pretenda ampliarlas al máximo, cuando queremos trazar las ambiciosas fronteras de lo físico que desea el fisicalismo es exigible el cumplimiento de ciertas condiciones.

Presentaré a continuación cinco condiciones que creo necesarias y suficientes para poder trazar las fronteras de lo físico (y señalaré algunas analogías ilustrativas respecto al caso de los países y sus fronteras):

1) Condición existencial: Ha de existir siempre algo que pueda reconocerse plausiblemente como físico. Sin esto, la tesis fisicalista sería literalmente falsa.

(Analogía con las fronteras de los países: Simplemente, sin fronteras no puede haber fronteras que ampliar. Una cosa es instaurar unas fronteras y otra, muy distinta, ampliarlas.)

2) Condición de continuidad: Podrán cambiar enormemente nuestras creencias y nuestras teorías físicas, incluso podremos acabar llegando a teorías inconmensurables con las que teníamos en un principio, pero deberá también existir siempre algo (una historia, una serie de referencias indécimas, etc.) capaz de garantizar explícitamente que esas creencias y teorías siguen siendo acerca de la realidad física.

(Aunque cambien completamente las fronteras de un país, debe haber siempre algo que garantice que se trata del mismo país y no de otro país diferente.)

3) Condición de contraste: Aunque la pretensión fisicalista consista en que todo lo que hay es físico, ha de haber siempre algo que directamente se resista a ser considerado físico. Sin esto, la tesis del fisicalismo sería vacuamente verdadera.

(La pretensión que un país tiene de ampliar sus fronteras al máximo sería vacua si no existieran territorios que se resisten a estar dentro de sus fronteras.)

4) Condición sobre las alternativas fisicalistas posibles: Frente a lo que se resiste a ser considerado físico, las alternativas fisicalistas posibles han de ser

1. Su reducción ontológica a algo físico.
2. Su inclusión en lo físico.
3. Su eliminación ontológica.

(Ante cierto territorio, un país que quiera ampliar sus fronteras tiene justo esas mismas opciones: reconocer que ya le pertenece, incluirlo dentro de sus fronteras o eliminarlo de las cartografías.)

5) Condición sobre los criterios de éxito y fracaso: El logro en ir reduciendo, incluyendo o eliminando todo aquello que ofrezca resistencia a ser

considerado físico ha de suponer el progresivo éxito de la pretensión fiscalista de que todo lo que hay es físico. Pero la existencia de algo que se resista definitivamente a ser considerado físico y que no pueda ser ni reducido a algo físico, ni incluido en lo físico ni eliminado ontológicamente de la realidad ha de suponer el fracaso de la pretensión fiscalista.

(Un país tendría progresivamente éxito en sus pretensiones de ampliar sus fronteras si, ante cualquier territorio, consigue reconocerlo como ya propio, incluirlo en sus fronteras o eliminarlo de las cartografías. Y tales pretensiones se verían frustradas si hay un sólo territorio que no pueda reconocer como propio, ni incluirlo en sus fronteras ni eliminarlo de las cartografías.)

En cierto sentido, el fiscalismo es "imperialista". Pretende ampliar sus fronteras hasta los límites de toda la realidad. Indudablemente, esa pretensión dejaría de tener sentido si de hecho sus fronteras llegaran a abarcar toda la realidad. Unas fronteras de ese tipo dejarían de ser fronteras. Pero sin llegar hasta ese punto, el fiscalista puede tener éxito en su pretensión de que todo lo que hay sea físico, o puede fracasar en tal pretensión. Y eso dependerá de su capacidad para ir reduciendo a lo físico, incluyendo en lo físico o eliminando de la realidad todas las cosas que se le resistan.

Aunque algunos matices sean importantes, lo que estoy proponiendo no es nada nuevo. Se trata, simplemente, de considerar al fiscalismo como una hipótesis empírica de gran generalidad. Es una hipótesis sumamente peculiar. Por ejemplo, sólo puede resultarnos interesante mientras su verdad no haya sido definitivamente establecida. Y bastaría un solo fracaso para rechazar tal hipótesis. Pero no sería más que una hipótesis empírica sometida a las cinco condiciones anteriores.

2. Maneras de construir lo físico

Hay muchas maneras de intentar construir el dominio de lo físico que serían incapaces de satisfacer esas condiciones. Veamos algunas de ellas.

1. Es físico todo aquello, y sólo aquello, que pueda ser descrito en un lenguaje fiscalista o cuyas descripciones sean traducibles a descripciones realizadas en un lenguaje fiscalista.

Se ha discutido mucho sobre esta manera de construir lo físico. Y ésta fué la opción seguida por algunos célebres fiscalistas. Pero, aún suponiendo que pueda especificarse adecuadamente un lenguaje fiscalista, que existan tales descripciones, etc., y que puedan satisfacerse las condiciones 1, 2 y 3, no tendríamos garantías de que se satisfacen las condiciones 4 y 5. La existencia de descripciones o traducciones fiscalistas es, por ejemplo, perfectamente compatible con un paralelismo ontológico.

2. Es físico todo aquello, y sólo aquello, que sea reconocido por nuestra física o que pueda ser reducido ontológicamente a algo reconocido por nuestra física.

También se ha discutido mucho esta posición. Como contrapartida ontológica de la posición 1, puede sortear problemas como el del paralelismo. Sin embargo, y aunque supongamos una satisfacción de 1, 2 y 3, aquí claramente no se satisfacerían las condiciones 4 y 5. La posibilidad de incluir en lo físico algo que no puede reducirse ni eliminarse no se contempla como una posibilidad fiscalista viable, ni entra en juego a la hora de evaluar el éxito o el fracaso de la tesis fiscalista.

3. Es físico todo aquello, y sólo aquello, que la física más desarrollada acabe reconociendo como físico.

Aún suponiendo que tenga sentido la noción de una física plenamente desarrollada (lo cual ha sido puesto en cuestión, con serios argumentos, por Crane y Mellor, 1990), y aún suponiendo que podamos interpretar ese "reconocimiento" como dando cabida a las alternativas de la condición 4, y permitiendo también una satisfacción de 5, esta manera de construir el dominio físico no garantizaría la satisfacción de nuestras condiciones 1, 2 y 3. Nada asegura que la física más desarrollada no acabe, por ejemplo, como la alquimia (con lo cual el fiscalismo podría ser literalmente falso). Nada asegura explícitamente que lo que esa física diga se refiera a la realidad física (a pesar del nombre). Y nada asegura que haya siempre algo que se resista directamente a ser considerado físico (con lo cual el fiscalismo podría ser vacuamente verdadero).

4. Es físico todo aquello, y sólo aquello, que quedaría implicado por una explicación fiscalista plenamente desarrollada del mundo.

Esta posición ha sido recientemente defendida por Jackson (1998: cap.1). Está muy cercana a 3 y tiene ventajas sobre ella. Jackson disipa, por ejemplo, gran parte de las dudas acerca del sentido que puede tener hablar de una "explicación fiscalista plenamente desarrollada del mundo". Ese sentido podría perfilarse mediante definiciones ostensivas de cosas físicas que debería reconocer cualquier física del futuro, mediante una especificación de los tipos relevantes de propiedades que en cualquier caso deberían contar como físicas y mediante el uso continuado de micro-explicaciones (véase Jackson, 1998: 6-8. Sobre el último punto, véase asimismo Pettit, 1993). Esto podría bastar para garantizar la satisfacción de nuestra condición 2. Sin embargo, y suponiendo que también se consigan satisfacer las condiciones 4 y 5, sigue sin quedar garantizada la satisfacción de las condiciones 1 y 3. Y por las mismas razones que en la posición 3.

3. La imagen estratificada de la realidad

Hasta el momento, no ha sido necesario distinguir entre entidades físicas y propiedades físicas. Lo haremos ahora examinando otra manera de construir el dominio de lo físico. Se trata de lo que habitualmente se conoce como imagen estratificada de la realidad.

Esta propuesta fue presentada en detalle a mediados del presente siglo por Putnam y Oppenheim (véase, por ejemplo, Oppenheim y Putnam, 1958), y ha sido revitalizada muy recientemente por Jaegwon Kim (véase, por ejemplo, Kim, 1993 y 1998. Véase también Pérez y Sabatés, 1995). Aunque abundan las lecturas pluralistas y las lecturas emergentistas de tal imagen, creo que su espíritu es fuertemente fisicalista. Y así es como voy a interpretarla aquí.

Recogeré algunos de los puntos más importantes de tal imagen estratificada de la realidad basándome en Kim:

1. Son físicas las entidades y propiedades reconocidas como tales por nuestra ciencia física.
2. Es una entidad física cualquier agregado de entidades físicas.
3. Son físicas las (macro-)propiedades de los agregados que se definan en términos de (micro-)propiedades físicas de sus partes físicas.
4. Es física cualquier propiedad de segundo orden definida sobre otras propiedades físicas.
5. Ninguna otra supuesta entidad ni propiedad puede ser real a no ser que sea reducible a alguna entidad o propiedad física de los anteriores tipos.

Lo que obtenemos es una construcción progresiva del dominio físico a partir de 1. Las entidades y propiedades mencionadas en 1 constituirían la base constructiva. Sobre esa base, se iría ampliando el dominio de entidades, con 2, y de propiedades, con 3 y 4. Las tesis 2, 3 y 4 constituirían los principios de clausura del dominio de entidades y propiedades físicas. Los agregados de entidades físicas dan lugar a nuevas entidades físicas ocupando diferentes niveles o estratos de la realidad. Y tanto las definiciones de (macro-)propiedades en términos de (micro-)propiedades como las definiciones de propiedades de segundo orden a través de una cuantificación sobre otras propiedades sirven para introducir nuevas propiedades físicas. El segundo procedimiento permitiría construir diferentes órdenes de propiedades dentro de cada nivel. A su vez, 5 es un principio de cierre. La reducción o la eliminación son aquí las únicas alternativas frente a cualquier supuesta entidad o propiedad no directamente reconocida como física.

Sin duda, lo que acabo de decir resulta insuficiente. Y quedarán muchas cuestiones pendientes. Pero creo que bastaría para preguntarnos si esta imagen es capaz de satisfacer nuestras cinco condiciones anteriores.

Supongamos satisfecha la condición de contraste 3. Es decir, supongamos que el fisicalismo expresado en la imagen estratificada de la realidad no es vacuamente verdadero. ¿Qué hay de las restantes condiciones? Mi impresión es que esta manera de construir el dominio físico tiene serios problemas para garantizar una satisfacción adecuada.

Respecto a las condiciones 1 y 2, nos encontraríamos justo en la misma situación que con las posiciones 3 y 4 examinadas en la sección anterior. Nada

asegura que la física no acabe, por ejemplo, como la alquimia, y que el conjunto de entidades y propiedades físicas básicas no sea simplemente el conjunto vacío (con lo cual, todos los niveles y órdenes de la imagen estratificada se vendrían abajo y el fisicalismo sería literalmente falso). Y nada asegura explícitamente que lo que la física diga se refiera a la realidad física (a pesar del nombre).

Y de una forma parecida a lo que ocurría con la anterior posición 2, la imagen estratificada de la realidad tampoco satisfacería las condiciones 4 y 5. Esta cuestión necesitaría muchos más detalles, pero el punto crucial sería el siguiente. El principio de clausura de propiedades basado en la construcción de propiedades de segundo orden está realmente lleno de problemas. Y ni una apelación a la noción de "superveniencia", ni otras estrategias, permite solucionarlos satisfactoriamente. El propio Kim ha renunciado recientemente a este principio de construcción de propiedades (véase Kim, 1993, y sobre todo Kim, 1998). Las propiedades de segundo orden no serían en realidad, para Kim, sino conceptos, o propiedades sólo en un sentido "abundante" o "latitudinario". Y deberían ser reducidas o eliminadas. No cabría otra opción fisicalista. El caso es que, si esto es sí, entonces la posibilidad de incluir en lo físico algo que no pueda ni reducirse, ni eliminarse, ni tampoco construirse de la manera como se construyen las (macro-)propiedades basadas en (micro-)propiedades, no se contempla como una posibilidad fisicalista viable, ni entra en juego a la hora de evaluar el éxito o el fracaso de la tesis fisicalista. Con lo cual, la imagen estratificada no será capaz de satisfacer nuestras condiciones 4 y 5.

4. Trazando las fronteras de lo físico

¿Qué podríamos hacer para trazar adecuadamente las fronteras de lo físico? ¿Hay alguna manera de construir el dominio físico capaz de satisfacer todas nuestras cinco condiciones?

Creo que una vez vistas las dificultades surgidas en otras posiciones, podríamos intentar algún tipo de respuesta estas preguntas. A fin de satisfacer las condiciones 1 y 2, parece necesario recurrir a algo, en cierto sentido, externo a nuestras creencias y teorías sobre lo físico, a algo que pueda garantizar la existencia de alguna clase de entidades y propiedades físicas y garantizar también una continua referencia a lo físico. Y eso podemos tenerlo a través de ciertas ostensiones (justo como es sugerido por Jackson, 1998:6-8).

Gran parte de la realidad que nos rodea debe ser física sean cuales sean nuestras creencias y teorías sobre el mundo físico. Cambien como cambien nuestras teorías físicas, deberán contar como entidades y propiedades físicas las mesas, sillas, el humo de los coches, las piedras con su peso y su dureza, los relojes, etc. Todo esto pertenece a nuestro mundo ordinario. Pero aquí será donde tendremos que buscar las ostensiones apropiadas. Y si nos olvidamos

de estas referencias ostensivas a lo físico, resultará muy difícil satisfacer las condiciones 1 y 2.

Paralelamente a esto, habrá también siempre otras muchas entidades y propiedades de nuestro mundo ordinario que se resisten a ser consideradas físicas. Nuestros pensamientos y esperanzas, nuestros dolores y nuestros miedos, por ejemplo. Y con ello, la satisfacción de la condición 3 estará más que asegurada.

¿Qué hay de las condiciones 4 y 5? Creo que aquí es decisivo proponer criterios precisos de inclusión.

La construcción de (macro-)propiedades físicas a través de (micro-)propiedades de las partes puede seguir siendo aceptable. El problema es que parece haber muchas (macro-)propiedades que no son fácilmente caracterizables de esta forma. Pensemos simplemente en las propiedades mentales. O, más en general, en todas las propiedades funcionales. ¿Qué hacer con ellas? Sin otros criterios de inclusión, y supuesta su no reducción a lo físico, las únicas opciones serían su eliminación o el fracaso de las pretensiones fisicalistas.

La construcción de propiedades físicas mediante su clausura respecto a operaciones booleanas tal vez pudiera servir. Las propiedades funcionales podrían acaso definirse mediante disyunciones infinitas de propiedades físicas. Pero esta estrategia introduce nuevos problemas de difícil tratamiento. Abre, por ejemplo, la puerta a multitud de meras propiedades tipo "Cambridge". Realmente, la mayoría de los autores prefieren considerar que el conjunto de propiedades físicas no está cerrado bajo las operaciones booleanas.

¿A qué otros criterios apelar para justificar la inclusión de una propiedad en el dominio físico? Creo que en este punto podría utilizarse muy provechosamente algo como el principio de clausura causal de lo físico. De acuerdo a este principio, podríamos formular el siguiente criterio de inclusión en lo físico:

Ha de ser considerada física toda propiedad tal que al ejemplificarse se produzcan efectos pertenecientes a una determinada clase física de eventos tipificada por nuestros conceptos físicos.

Haciendo uso de este criterio, podremos llegar a considerar físicas muchas propiedades resistentes en principio a esta consideración. Podríamos llegar a considerarlas físicas cuando son ejemplificadas en ciertas clases de objetos y en ciertas clases de circunstancias. Muchas propiedades mentales, funcionales, computacionales, etc., adquirirían el estatus de propiedades físicas en relación a ciertos objetos y circunstancias de ejemplificación. Y perderían tal estatus en relación a otros objetos y circunstancias de ejemplificación.

Y bastarían ciertas ostensiones apropiadas, junto con esta variedad de conexión causal, para poder ampliar las fronteras de lo físico de manera

inclusiva. Sin que fuera necesario optar por el eliminativismo o por el fracaso de la pretensión fisicalista.

Mientras podamos mostrar, aunque sólo sea ostensivamente, propiedades conectadas causalmente con clases físicas de eventos tipificados por nuestros conceptos físicos, habrá inclusión. Y el fisicalismo será un éxito a pesar de no darse ninguna reducción ni eliminación. Y mientras podamos también mostrar propiedades que se resistan a esa conexión causal, ese éxito no supondrá una verdad vacua.

Las propiedades que, en ciertas circunstancias y respecto a ciertos objetos, sean capaces de ser incluidas de esta manera en lo físico pueden ser las mismas propiedades que, en otras circunstancias o respecto a otros objetos, se resistan a tal inclusión. Y en este segundo caso, un fisicalista sí debería optar por el eliminativismo. Pero no en el primero.

De cualquier forma, una vez adoptado el anterior criterio de inclusión, las fronteras del dominio de propiedades físicas acaban siendo algo muy relativo a las condiciones de ejemplificación de las propiedades. Y esto conduciría a nuevas formas de entender la doctrina fisicalista que merecería la pena explorar.

5. Reflexiones finales

Al principio del trabajo, me referí brevemente al desprestigio que suelen tener hoy día posiciones metafísicas tan ambiciosas como el fisicalismo. La metáfora de las fronteras me ha servido para mostrar que la ambición fisicalista puede ser formulada coherentemente. Y que la razón de tal desprestigio no puede ser que el fisicalismo esté obligado a ser vacuamente verdadero o literalmente falso.

Ciertamente, el uso que he hecho de esa metáfora de las fronteras sugiere un impulso, digamos, "imperialista" por parte del fisicalismo. He querido enfatizar a propósito este rasgo. Y es hora de preguntarse qué puede haber de malo en tal impulso.

Creo que en la pretensión fisicalista acerca de que todo lo que hay es físico no hay más maldad que la maldad general que pueda haber en la pretensión de que sea verdad que p cuando afirmo que p (para cualquier p). Sí hay una diferencia en el contenido. El contenido de la pretensión fisicalista afectaría a todo lo que hay. Su verdad afectaría a todo. Pero ver aquí alguna maldad intrínseca no sería ni más ni menos que ver una maldad intrínseca en algunas verdades.

BIBLIOGRAFÍA

- Crane, T. and D.H. Mellor, "There is no question of physicalism", *Mind*, 99, 1990 [reprinted with a postscript in P. Moser and J.D. Trout (eds.) *Contemporary Materialism*, London, Routledge, 1995].
- Esquivel, J., *La polémica del materialismo*, Madrid, Tecnos, 1982.

Jackson, F., *From Metaphysics to Ethics. A Defence of Conceptual Analysis*, Oxford, Clarendon Press, 1998.

Kim, J., *Supervenience and Mind*, Cambridge, Cambridge Univ. Press, 1993.

-----, *Mind in a Physical World*, Cambridge, MIT Press, 1998.

Moulines, C.U., *Exploraciones metacientíficas*, Madrid, Alianza, 1982.

Oppenheim, P. and H. Putnam, "Unity of Science as a Working Hypothesis", in H. Feigl et al., *Minnesota Studies in the Philosophy of Science*, vol. II, Minneapolis, Univ. of Minnesota Press., 1958.

Pérez, D., y M. Sabatés, "La noción de superveniencia en la visión estratificada del mundo", *Análisis Filosófico*, vol. XV, 1995, 181-199.

Pettit, P., "A Definition of Physicalism", *Analysis*, 53, 1993, 213-23.

Manuel Liz

Univ. de La Laguna

e-mail: manuliz@ull.es





LAWS, DEPENDENCE AND FORMAL RELATIONS

JOAN PAGÈS

Universitat de Girona

ABSTRACT: The Inference Problem is an objection posed against some theories of natural laws such as those advanced by Dretske, Tooley and Armstrong, which conceive them as relations between universals. I will expose my own solution to the Inference Problem by assuming an immanentistic theory of universals and laws. I intend my solution to be compatible with any version of the Dretske-Tooley-Armstrong theory which is at least committed to a local version of immanentism and to a certain formal relation holding between particular states of affairs.

KEYWORDS: Laws, universals, dependence.

The Inference Problem is an objection posed by different authors to some theories of natural laws as those advanced by Dretske (1977), Tooley (1977, 1987) and Armstrong (1983) which conceive them as relations between universals (called "DTA theories" henceforth). Notice that it is a basic, commonsense fact that laws imply their corresponding regularities; no more than conceptual knowledge is needed to infer (1) 'All Fs are Gs' from (2) 'It is a law that all Fs are Gs'. This Moorean fact is usually taken as a correction criterion for theories of natural laws. Then, the problem for the DTA theorist consists, firstly, in showing that this inference can be *prima facie* warranted given that laws are relations between universals and, secondly, in giving an explanation of the way in which laws so conceived imply their corresponding regularities. As many DTA critics have claimed, the attempts of DTA theorists to accomplish the two tasks we have just described are far from satisfactory. A solution by postulate leaves the nature of the inference unexplained, and simply begs the question raised by the first task. Other apparently different attempts do not work, for a variety of reasons. Armstrong's answer to Van Fraassen's version of the problem (Armstrong 1993, Van Fraassen 1989), Tooley's "speculative solution" or even Sider's proposals on Tooley's line (Tooley 1987, Sider 1992) are doomed to failure (Van Fraassen 1993, Carroll 1994, Sider 1992). In addition, suggestions for a solution to the inference problem from a trope theory of laws, such as those suggested by Fuhrman (1991), also seem to be unsuccessful, for two main reasons: its commitment to *possibilia*, and, more particularly, its rigid pattern, which impedes its extension to cover

exclusion laws or cases involving regularities whose conditional contains an antecedent and a consequent involving different particulars.

In my contribution to this meeting I will expound my own solution to the Inference Problem, showing how the two tasks mentioned above can be accomplished from an immanentistic theory of universals and laws. I intend my solution to be compatible with any version of the DTA theory which is at least committed to a local version of immanentism (which holds that universals are a kind of dependent entity whose existence depends on the existence of a particular that exemplifies them, and also, more specifically, that universals have spatio-temporal location: they are completely located wherever their instances are) and (at least for some cases) to certain formal relations holding between particular states of affairs.

In the first place, there is no special difficulty in showing how, in principle, regularities could be inferred from the corresponding nomic facts that the DTA theorist postulates. Reflection makes it evident that some traits can be analytically inferred from determinables to their determinates (and conversely). Note, for instance, that the following inference is analytically valid:

(I) *a* is blue

a is coloured.

(The following completing premisses allow us to logically infer (I)'s conclusion from (I)'s premiss: (a) Blue is a colour; (b) For any object *x*, *x* is coloured iff *x* has a property which has the property of being a colour.)

Similarly, the following inference is also analytically valid:

(II)

Magenta is a red

Red is a colour

Magenta is a colour

A Dependence relation

Let us now say that a set of properties, *D*, is a set of determinates for a certain determinable property, *P*, if and only if the following conditions are fulfilled:

(i) Necessarily, if any particular instantiates any property belonging to *D*, then it also instantiates the determinable *P*.

(ii) Necessarily, if any particular instantiates the determinable *P*, then it also instantiates one and only one property belonging to *D*¹.

The set of colours and the property of being coloured, or the set of shapes and the property of having shape, satisfy these conditions. The first condition requires that if *a* is, let us say, red, then *a* is coloured, whereas the second condition has it that if any particular object is coloured, then it has one and only one particular colour². In the light of condition (ii), these

¹ These conditions are Johnson's old axioms (Johnson 1921,1922). Another classic study of determinate/determinable relations is Prior's (Prior 1949).

² There is a slight difficulty here relative to condition (ii), for it is obvious that in a certain sense an object can be multi-coloured. For the sake of simplicity, the domain of objects must be restricted to objects which, being coloured, are chromatically uniform, or

considerations should be reformulated. Since one and the same coloured particular can be red and magenta at the same time, the set of colours does not constitute a set of determinates for the property of being coloured. Red and magenta are not, however, at the same level of determination; there is a set of determinates for the determinable red to which magenta belongs. Hence, red is a determinate for being coloured, but it is a determinable with respect to magenta, which is also a determinate for being coloured. Then, instead of focusing on the set of colours, as we have just done for the sake of simplicity, it would probably be more accurate to focus on any one of its subsets, all of whose properties are at the same level of determination.

It is also interesting to consider whether the trait transferrings from determinables to their determinates and from universals to their instantiations run together. The following considerations are intended to show that this is the case under the assumption of a certain sort of immanentism. Granted, it cannot be said that universals and their instantiations satisfy axioms (i) and (ii), for it is assumed in these axioms that they can be satisfied only by instantiable entities, but instantiations are not instantiable entities. However, we shall see next that at a more abstract level a close link between the two sorts of entity pair can be found.

Let us assume that universals are immanent entities in the strong, local sense that they are located wherever their instances are. Note in the first place that axioms (i) and (ii), which describe the relations between determinates and determinables, can be rewritten as follows:

(DET) A property, *P*, is a determinable for a set of properties, *D*, if and only if the following conditions are satisfied:

(i) Necessarily: for any spatio-temporal region, *s*, if some property in *D* exists at *s*, then *P* exists at *s*,

(ii) Necessarily: for any spatio-temporal region, *s*, if *P* exists at *s*, then there is one and only one property in *D* that exists at *s*.

Observe also that (DET) falls under the following pattern:

(DEP) For any entity, *f*, and class of entities, *E*, *f* is related by *R* to any of the entities of class *E* iff:

(i) Necessarily: for any spatio-temporal region, *s*, if any *e* in *E* exists in *s* throughout, then *f* exists in *s* throughout,

(ii) Necessarily: for any spatio-temporal region, *s*, if *f* exists in *s* throughout, then there is one and only one *e* in *E* that exists in *s* throughout.

We call the last definition (DEP) because it just describes a dependence relation, *R*, holding between the entities that satisfy it. It should be noted that if the local immanentistic conception of universals is to be accepted then any universal is related by *R* to any of its exemplifications, as any determinable is to any of its determinates. In fact, this would be an insightful way to specify the local immanentistic conception of universals. It should be understood, then, that determinable/determinate and universal/instantiation pairs have something in common: both are related by the same kind of dependence relation. This connection allows us to hold that if transmissions from determinables to

better perhaps, we should understand 'being of a certain colour' as meaning being completely of a certain colour. The uniqueness requirement merely intends to preserve the obvious fact that no object can be completely red and completely blue simultaneously.

determinates of those traits which are dependent on the dependence relation R - which we know holds between determinables and their determinates - are acceptable, then the transmission of traits from universals to their instantiations should also be acceptable, granted that these traits are dependent on the dependence relation R which also holds between these latter entities³. Although we are not yet in a position to argue conclusively that this option can be successfully developed, we shall give a rough sketch just to indicate that the chances of doing so are good. Perhaps not all traits of universals can be transferred to their instantiations, but the relational trait of necessitation is transferable because it is a trait partially analyzable in terms of existential dependence, so it will be transferred from universals to their instantiations precisely in virtue of the relations of existential dependence holding between these entities, relations which are explicitly described by (DEP). The next section of this paper is devoted to developing this idea in some detail.

Notice also that the uniqueness requirement embedded in (DEP)'s condition (ii) assumes the principle for concrete particulars, (C), according to which no two different concrete particulars can occupy the same spatio-temporal region. More precisely:

(C) For any particulars, x and y , if for any spatio-temporal region r , (x exists at r if and only if y exists at r), then $x = y$.

Though this principle is quite controversial, as we will see, my solution to the inference problem does not rely on it.

Standard Law-Statements

The ingredients needed to solve the inference problem are the following:

(a) The thesis, (D), holding that the relation between universals and their instantiations satisfies (DEP), which presupposes the local immanentistic thesis for universals:

(LI) For any possible world, w , and any n -adic universal, U , U exists at w if and only if at w : there is a spatio-temporal region, r , such that there exist n particulars at r , $x_1, \dots, x_n$, such that at w : $x_1, \dots, x_n$ exemplify (in this order) U at r . Then it is said that U (at least partially) exists at r at that world w .

(b) A formal relation, R , holding between particular states of affairs, defined as follows. For any two spatio-temporal regions r and s and any two properties, F and G , $e_r(F)$ is related by R to $e_s(G)$ if and only if there is an x such that x is F at r and x is G at s , and x is the particular component of $e_r(F)$

³ Yablo defines a determination relation between properties that does not assume immanentism; in fact, it does not even assume that properties are universals (Yablo 1992). Yablo's definition stipulates that a property P determines a property Q if and only if (i) necessarily: for any x , if x has P then x has Q and (ii) it is possible that for some x , x has Q but not P . (Yablo 1992, p. 252.) On the other hand, Yablo also defines a determination relation between events that allows for certain pairs of events being in the determinate/determinable relation. The definition is as follows: an event p determines another event q if and only if: for p to occur (in a possible world) is for q to occur (in that possible world), not *simpliciter*, but in a certain way. For instance, Icarus flying too near the sun determines his flying. (See Yablo 1992, pp. 260-261).

and $e_S(G)$ (where $e_r(F)$ and $e_S(G)$ are the corresponding instantiations of universal F at r and universal G at s .)

(c) A partial analysis of the nomic necessitation relation, N , which is responsible of nomic facts in a DTA theory and which holds between two universals, as a relation of existential dependence between them as follows:

(Nec) If F is related to G by N , then for any spatio-temporal region, r , if F exists at r , then G exists at r .

(d) An analysis for law-statements:

It is a law that all F s are G s is true if and only if (a) $N(F,G)$, and (b) for any spatio-temporal region r such that $e_r(F)$ exists: if $e_r(G)$ exists then $R(e_r(F), e_r(G))$.

Some remarks are in order. (Nec) just tries to make explicit part of the intuitive content that a relation called “necessitation relation” is intended to have; in this sense, it is conceived as a relation of existential dependence that holds between the related terms, a pair of universals. It is precisely the fact that the necessitation relation entails an existential dependence that allows the use of (DEP) to license the inference, for (DEP) just describes existential dependences in a double route, from instantiations to universals and from universals to instantiations. Notice, moreover, that this partial definition does not include as a definitory trait the inference thesis, so it is not subject to the objections raised against Armstrong’s and Tooley’s attempts to solve the inference problem. It is true that the main reason for partially characterizing N by (Nec) is to preserve the inference, but this partial definition helps to explain how and why the inference is possible. In addition, no natural necessity operator is involved in this solution, and there is no need at all to postulate *possibilia*⁴. Finally, this solution does not assume any particular version of (strong) local immanentism such as Armstrong’s theory of laws as universal states of affairs though it is compatible with such a theory.

Let us see now how these ingredients allow us to solve the inference problem. Let us suppose that:

(1) F necessitates G ,

and also that:

(2) At s : a is F .

Now, given (2), by (D) and (DEP), (i):

(3) F exists at s .

Then, by (1) and (Nec):

(4) G exists at s .

Now, by (D) and (DEP), (ii):

(5) There is an instantiation of G , α , such that α exists at s .

Thus, by (b):

⁴ This is not committed to the idea that there is no need for modal components in the analysis (even partial) of the nomic necessitation relation. That is, it is perfectly possible that other traits will be used in its characterization which include modal notions with the purpose of covering some other aspects that laws are supposed to have. For instance, such a trait could be called for in the explanation of how laws support counterfactuals. For an argument that it is not necessary to insert modal notions in the analysis of the nomic relation to account for the counterfactual supporting problem, see my 1997.

(6) $R(a \text{ being } F \text{ at } s, \alpha)$,
 and by definition of R :

(7) a is G at s .

It should be observed that the uniqueness condition embedded in (DEP), (ii), would have allowed us to infer (7) without resorting to the properties of the formal relation, R . There are, however, two reasons that justify not to appeal to (C). A first, but weaker reason, is just the controversial character of (C) itself. A second and stronger reason is that formal relations are required at any case in dealing with laws implying other kind of regularities; for instances, those regularities involving two different instantiations of one and the same particular at different spatio-temporal regions.

Exclusion Law-Statements

Besides the advantages of this solution for the inference problem already mentioned, we should add the fact that it is easily extended to cover some other types of laws. Take, for instance, an exclusion law such as:

(EL) It is a law that all F s are not G .

A truthmaker for (EL) should include a nomic relation, different from the necessitation relation partially defined by (Nec); let us say an "exclusion relation", E . Now, in order to preserve the inference from (EL) to the corresponding regularity that all F s are not G s, we can partially define E as a relation that satisfies:

(Exc) If F is related to G by E , then: for any spatio-temporal region, r , if F exists at r then G does not exist at r .

With both this partial analysis of the nomic relation of exclusion and thesis (D) at hand it is an easy matter to derive from (EL) the corresponding regularity.

Let us assume that:

(1) F is related to G by E

and that:

(2) At s : a is F .

Then, by (D), (i):

(3) F exists at s .

By (Exc):

(4) G does not exist at s .

Now, by (4), (D) and (i) it follows that:

(5) At s nothing exists that it is an instantiation of G .

Hence,

(6) At s : a is not G .

In inferring that all F s are G s from the relational nomic fact involving the necessitation relation, we had to appeal to the properties of the formal relations R ; but note that this is not the case here. The resort to those properties is no longer necessary because of the nature of the regularity involved. Notice that the regularity associated to the necessitation relation states that for any particular instantiating the antecedent property that *same* particular *also* instantiates the consequent property. (Nec) only guarantees that the consequent property must be instantiated at exactly the same region as the antecedent property is instantiated, but this is not enough to warrant that the same particular is involved in both instantiations. In exclusion regularities –for all F , that F is not

G— there is still the same particular involved in the antecedent and the consequent, but the consequent makes no claim that that particular must exemplify the consequent property⁵.

Causal Law-Statements

On the other hand, other regularities can involve, in principle, different particulars located in different spatio-temporal regions. We are thinking of simple formal schemes of causal laws such as:

(LC) It is a (causal) law that for all x , if x is F then there is a y such that y is G .

In this case, inferences involving the corresponding regularities do not require the formal relation R as is the case when dealing with exclusion laws. Besides, unlike exclusion laws, (LC) does not need local immanentism in any version thereof; global immanentism is good enough. As we have characterized it, global immanentism holds that:

(GI) For any possible world, w , and any n -adic universal, U , U exists at w if and only if there exist n particulars $x_1, \dots, x_n$ at w such that $x_1, \dots, x_n$ exemplify (in this order) U at w .

Now we turn to kinds of law that include information about the spatio-temporal relations holding between the antecedent fact and the consequent fact. Consider, for example, the following simple schematic case:

(L) It is a law that for any particular x , if x is F at time t , then there is a particular, y , such that y is G at time $t + \delta$.

In this instance, it is part of the content of the law-statement that the consequent fact must occur after a certain temporal interval whose value is δ . This case can be subsumed under the following pattern of law statement:

(LC') It is a (causal) law that for any particular x , and any spatio-temporal region r , if x is F at r , then there is a particular, y , and a spatio-temporal region, s , such that y is G at s and $s = V(r)$.

(where 'V' designates a spatio-temporal variation function; that is, a function which takes spatio-temporal locations as arguments and assigns to them other spatio-temporal locations.)

To account for these laws we should incorporate the nomically fixed spatio-temporal variations into the nomic relations, specifying them in the partial definition. In fact, we will not define just one more sophisticated nomic relation, but a family of them, each member of the family being associated to a distinct variation function, V . Thus, for any variation function, V , we partially define its associated nomic necessitation relation as follows:

(Nec V) F is related to G by NV only if for any spatio-temporal region r : if F exists at region r , G exists at region $V(r)$.

It should now be obvious that (Nec V) and (D) jointly allow us to infer from (LC') the corresponding regularity. In this case, the resort to the properties to the formal relation R is unnecessary too. Notice also that this

⁵ We should bear in mind that we are presupposing that 'F' and 'G' designate genuine universals. General reasons for avoiding negative universals notwithstanding, there is no need at all to require them to account for the applying of the predicate 'no-G' to particulars under the assumption that G is a true universal.

device can be easily extended to successfully cover exclusion laws involving different locations.

Some concluding remarks are in order. We have seen how to solve the inference problem for all types of laws we have considered by postulating a nomic necessitation second order relation partially defined as a dependence relation between universals, a variation location function (which, as we saw, is required to account for laws involving spatio-temporal relations between particular states of affairs), a formal relation (which is required by laws involving properties instantiated by the same particular at the same location) and also by holding local immanentism.

Our solution seems overtly superior to those advocated by Armstrong, Tooley or Fuhrman, for no claim for *possibilia* is made, no *ad hoc* postulate is stated, the solution covers a fairly wide range of laws, and the inference from laws to regularities is not merely made possible but its nature successfully explained.

REFERENCES

- Armstrong, D. M.: *What is a Law of Nature?*, Cambridge Studies in Philosophy. Cambridge University Press, 1983.
- Armstrong, D. M.: "The Identification Problem and the Inference Problem", *Philosophical and Phenomenological Research*, 53 (2), 1993, pp. 421-422.
- Bergmann, G.: "Event Causation: The Counterfactual Analysis", *Philosophy and Phenomenological Research*, X, 2, 1949, pp. 263-264.
- Bigelow, J. and Pargetter, R.: *Science and Necessity*, Cambridge University Press, 1990.
- Carroll, J. W.: *Laws of Nature*, Cambridge University Press, 1994.
- Dretske, Fred: "Laws of Nature", *Philosophy of Science*, 44, 1977, pp. 248-268.
- Fuhrman, A.: "Tropes and Laws", *Philosophical Studies*, 63, 1991, pp. 57-82.
- Pagès, J.: "Armstrong on the Roles of Laws in Counterfactual Supporting", *Theoria*, 12, 29, 1997, pp. 337-342.
- Tooley, M.: "The Nature of Law", *Canadian Journal of Philosophy*, 7, 1977, pp. 667-698.
- Tooley, M.: *Causation. A Realist Approach*, Clarendon Press, Oxford University Press, 1987.
- Van Fraassen, B.: *Laws and Symmetry*, Clarendon Paperbacks, Oxford University Press, 1989.
- Van Fraassen, B.: "Armstrong, Cartwright and Earman on 'Laws and Symmetry'", *Philosophical and Phenomenological Research*, 53 (2), 1993, pp. 431-444.

Joan Pagès
 Universidad de Girona
 e-mail: pages@skywalker.udg.es

THE ONTOLOGY OF NORMS AND THE DECLINE OF THE ANALYTICAL LEGAL POSITIVISM

LUCA PARISOLI

Centre de Théorie du Droit, Université Paris X

ABSTRACT: According to some analytical philosophers the concept of natural law is able to explain our legal systems without denying the irreducibility of our moral experiences. If we use only a poor ontology which denies any modality of existence to the object "natural norms" (instead of using a more rich ontology which admits some modalities of existence to the object "natural law") we cannot in fact be able to give a good account of our legal systems. Natural norms exist as well as positive norms. In fact, if we don't accept the existence of some immutable norms, then we cannot avoid that the legal system becomes an arbitrary system.

KEYWORDS: Natural law, super-constitutional norms, foundations of norms, juridical realism.

Les droits de l'homme sont aussi juridiques que tout autre section du droit. Lorsqu'une méthodologie se déclare incapable d'expliquer leur jeu, loin de les anathématiser, elle ne fait que reconnaître son impuissance à rendre compte de la réalité juridique

Andrés Ollero-Tassara, *Droit "positif" et droits de l'homme*, 1997, p. 76, spanish text in *Derechos humanos y metodologia juridica*, 1989, p. 123

I. Introduction

The contemporary discussion about the doctrine of natural law is constructed around many arguments which try to show that the same appeal to the concept of natural law is not justified. Nevertheless, this type of background (built on the theories of Kelsen) is now no more generally accepted. The analytical legal positivism dominium is in fact now menaced by some other analytical positions which try to demonstrate the utility of the same concept of natural law. According to some analytical philosophers, in fact, this concept is able to explain our legal systems without denying the irreducibility of our moral experiences: Xavier Dijon, for example, has used this concept in

order better to explain the same positive law¹; John Finnis has recently published a work about Saint Thomas and his political and juridical thought²; Francesco Viola has tried to overcome the facts/values gap³. In a certain way, we can say that now many legal naturalists try to use the history in order to justify their position and so they are able to follow the indications of Robin G. Collingwood according to which metaphysical propositions are also historical propositions⁴. But in another way all these scholars (Finnis, Djon, Viola) are perfectly able to be analytical philosophers so that they can show that the binomial legal-positivism/analytical-style has only been a simple contingency: today in fact it is possible to have a real analytical school of natural jurisprudence.

In this sense, the aim of my paper is either to point out some limits of a positivistic position, or to explain why the notion of natural law can be a useful notion: if we use only a poor ontology which denies any modality of existence to the object "natural norms" (instead of using a more rich ontology which admits an existence modality to the object "natural law") we cannot in fact be able to give a good account of our legal systems. Natural norms exist as well as positive norms and we can deny their existence only if we are concerned with an ideological operation which is not really able to hold fast to the analytical style.

In fact, if we don't accept the existence of some immutable norms, then we cannot avoid that legal system becomes an arbitrary system. In some way, the positivistic position is not able to avoid this consequence, because it cannot admit that these immutable norms can have the same existential statute of the modifiable norms. Nevertheless, immutable norms exist and they are norms which belong to the natural law. In this sense, today it is not only possible, but also desirable that many thinkers are analytical philosophers of natural jurisprudence (as Finnis, Djon and Viola are).

II. Super-constitutional norms

The core of my position is that a poor ontology, that is to say an ontology that denies the existence of "natural norms" as object of the world, is not an ontology able to give an account of the legal system as a rich ontology (according to which "natural norms" are an object of the world) is. Nevertheless, when I speak of the existence of natural norms as objects in the world, I do not want to say that they exist in time and space: there isn't any special "realm" where we can touch natural norms. In this sense, if we accept a radical physicalist gnoseology (where reality is reduced to empirical protocols), reality is just something whose substance and form may be perceived. As Georges Kalinowski has clearly shown⁵, in fact, this type of ontology does not allow us to conceive normative facts as "objects" of our world. But we are not obliged to accept such a poor ontology (Kalinowski himself has never accepted such an ontology and has rather constructed a neo-thomistic philosophical

¹ X. Djon, *Droit naturel*, I, Paris 1998.

² J. Finnis, *Aquinas*, Oxford 1998.

³ F. Viola, *Dalla natura ai diritti*, Roma 1995.

⁴ R.G. Collingwood, *An Essay on Metaphysics*, Oxford 1940.

⁵ G. Kalinowski, *L'impossible métaphysique*, Paris 1981.

system). Now, I do not want to propose any ambitious ontological program. Nevertheless, I would like to show that natural norms are an object of human thought. But they are not just the product of our intellect or of our will: they "are" before the mind. They are in fact the result of the discovery of what exists, as well as all mathematical theorems. According to the standard conception of mathematicians, in fact, theorems are not just a product of the intellect (even if we have to recognise that there are some nominalist mathematicians according to whom this thesis is not an acceptable one). The problem is that my position, as well as some others, is not universally accepted in that it cannot be really demonstrated: it can only be shown, as well as it can be accepted because of some consequences we have if we accept it.

If we deny the existence of natural norms, we are committing with an ideological position which I cannot accept because of many reasons which I am going to explain. In this sense, I use the expression "ideology" in a particular way, that is to say in the Marxist sense according to which an ideology is a theory, that it is also a theory only apparently interested in what it is true and what it is false: his real aim, in fact, is to influence actions and behaviours. In particular, I can call the positivist position an ideological one because according to many positivistic thinkers the master of law is the legislator: there is no bad or good statute (as a natural law philosophers can think), there is only a valid or invalid statute. Each legislator, as a new God, is the measure of all things.

In this sense I do not use the expression "ideology" as a way of criticising a "philosophical sentiment" which is different from mine. Bradley's position, for instance, according to which all relations are internal relations, is not an ideological position, even if we can think that it is really extravagant and even unacceptable. And analogously we can hold that Berkeley's idea that "esse est percipi" is not a correct theory, but we can never hold that Berkeley has an ideological position. Bradley's position, in fact, as well as Berkeley's one, has not the intention to influence human action as many ideological positions do even when they use technical instruments (juridical procedures) and try to be neutral and rigorous (pure science of law).

Coming back to the reasons I would like to analyse in order to refute the ideological positions of those who deny the existence of natural norms, I would like to begin with the analysis of a recent paper written by Michel Troper ("Les normes supraconstitutionnelles", translated in Italian as "Le norme sovraconstituzionali"⁶) who is a positivist and legal realist thinker, even if he does not accept any radical and positivistic thesis. In this article Troper holds that the idea of super-constitutional norms cannot be used as a positivistic argument because it has always been used by the philosophers of natural law. In particular, he points out that norms are the result of the interpretative activity of normative sources. But if this is true, then also super-constitutional norms must be the result of an interpretation of some source. Nevertheless, in this case we cannot find any source in our actual juridical systems: everywhere, in fact, Constitution is the highest legal source. But if we speak of super-constitutional norms, the legal source cannot obviously be the Constitution. Finally, then, the only source for super-constitutional norms may be natural law. On the other hand, if we do not think that there exists any natural law,

⁶ M. Troper, "Le norme sovraconstituzionali", *Analisi e diritto* 1996, pp. 255-274.

super-constitutional norms are only a *flatus vocis* without existence - and this is the solution which Troper arrives at.

For this reason I completely agree with the logical and rhetorical analysis of Troper about super-constitutional norms, but I cannot accept his conclusions. Troper, in fact, as positivist thinker, cannot accept the idea that the *prima facie* existence of super-constitutional norms is incompatible with a positivist position and then he proposes to ignore them: they are dangerous for a positivistic position, then we must no more speak of them. On the other hand, my opinion is that the existence of super-constitutional norms allow us to discard any form of positivism. I cannot in fact think that a philosopher of law's position could really be blind to the activity of the constitutional judge who usually uses super-constitutional norms and then usually accepts the existence of natural norms.

III. The symbolic value of norms

I know that my position is not a completely neutral one (may be that it is also an ideological position ?!), as well as I know that I accept the so called juridical and non-empiricist realism. Nevertheless, I would like to defend this juridical realism because I am convinced that any forms of meta-theoretical and absolutely neutral position is today no more *à la page*, as well as it is no more *à la page* any form of meta-ethics as it was conceived in the '50th. At the same time I am conscious that I cannot now construct a reasonable and personal theory of law, able to be clear and analytical as many positivist (but also non-positivist) theories. For this reason I would like only to point out some aspect of this theory, hoping that some others could accept the idea of constructing a new way of making normative propositions in the philosophy of law.

The first step is to assume the idea that Law is the product of the judge's activity when he attributes meanings to normative utterances in order to arrive at having norms. In this sense, when I speak of Law, I am speaking of the actual and present Law, which is completely different from the positive and statutory law which is rather constituted by all normative utterances established by competent authorities. We need in fact to make a clear distinction between the juridical sphere of norms and the enunciative sphere of normative utterances. But, on the other hand, we need to point out that one of the most important aims of a philosopher of law is to explain the existence of norms. In this sense the philosophy of law is different from the jurisprudence: jurisprudence, in fact, is made by judges who are political players and whose aim is to administrate justice; philosophy of law is made by philosophers who can try (and according to me have to) to argue in favour of a (rich or poor) ontology of norms. In this sense, from a philosophical point of view, the most important problem to analyse is that of the foundation of norms, even if according to many positivist philosophers, the problem of the foundation of norms is not a real problem to deal with. But according to positivist thinkers, the philosophy of law is just a descriptive science (as botanical science).

Nevertheless, the problem of the foundation of norms is extremely important if we want to prove that norms have a symbolic value - as Pierre Legendre has very clearly expressed by saying that norms without a symbolic

value are no more norms⁷-. And if we want to analyse the problem of the foundations of norms we have to say something more about super-constitutional norms. As I have just said, in fact, we can justify these norms only if we postulate the existence of natural law. In this sense, the existence of natural norms may be accepted also by those sceptical philosophers who cannot understand how it is possible to arrive at knowing which sort of norms they are, once given their existence. But I cannot now analyse also the gnoseological problem about natural law. Even if it is not impossible to arrive at an explanation (we can choose the explanation of Pufendorf, or that of Saint Thomas, or even that of Duns Scotus), this type of analysis is not different from that about moral values or psychological behaviours and for this reason it is an analysis which I could never exhaust in a short paper, even if I can suggest the Common sense philosophy of Scottish philosophers in XVIIIth century as a good starting point.

For this reason I come back to the existence of natural laws in order better to explain my convictions. So, after having pointed out the necessity of natural norms in order to justify the existence of super-constitutional norms (also with the involuntary aid of Michel Troper), I now need to explain the reason why we need super-constitutional norms. The existence of super-constitutional norms is in fact necessary in order to justify the fact that Constitutions is a normative document constituted by norms which are different from and superior to all other norms. In fact, constitutional norms are not modifiable by an usual act of legislation as ordinary norms are, as we have been shown by the traditional and constitutionalist political philosophy (from the Medieval Scholastic to Locke, from Wolff to Burke and the contemporary debates). Traditionally, in fact, Constitution is explained as a normative sphere able to limit the power of the sovereign. But this classical thesis is no more comprehensible if we accept the positivistic thesis according to which the sovereign has always the power to modify the Constitution as simple set of normative utterances. How can we in fact limit the power of the sovereign if he can modify the same Constitution?

We could of course decide not to analyse this type of problems: we could in fact decide that these problems are just political problems. But in this way we can only arrive at a conception of legal system completely disjoined from the real world of our experience.

On the other hand, if we want to arrive at a conception of legal system linked to the real world, we need an analysis of these problems. And the first step is the consciousness that the positivist analysis of Constitution as a set of normative utterances is not able to explain the existence of immutable norms which can never be modified by the simple decision of a legislator. A positivist account of Constitution as a simple set of normative utterance, in fact, is not able to give an account of the symbolic meaning that Constitution has always had and which has often been recognised by the use of expressions like "national identity", "document elaborated by the Founder Fathers", "Constitution of such a year", etc.⁸.

⁷ P. Legendre, *L'amour du censeur*, Paris 1974.

⁸ Cf. all the scientific research by Pierre Legendre, and very recently *Sur la question dogmatique en Occident*, Paris 1999.

IV. Towards a realistic philosophy of natural law

Any legislative formalism is able to warrant the legitimate existence of a good legislator who respects the immutable norms: we haven't any positivist justification of the illegitimacy of the action of a legislator who creates a normative utterance which is not compatible with an immutable norm.

Nevertheless, immutable norms exist, even if they are not universally accepted. And they exist either if there is a non-human authority able to found them, or if the belief in their existence is able to satisfy some criteria of belief's justification⁹. And if we analyse the constitutional jurisprudence of many countries, we can have a proof of what I have said. In fact, we can easily find some motivations which express the existence of immutable constitutional norms because there exist some super-constitutional norms able to warrant their immutability. The judge cannot in fact speak of immutable norms without admitting super-constitutional norms. But by admitting super-constitutional norms, the judges can not only found on this super-constitutional sphere his immutable norms, but they can also avoid the possibility that a legislator could limit their applications by means of new normative utterances which can always be modified.

Obviously, all this argument is linked to a particular conception of the activity of the judge. As I have just pointed out, in fact, the role of the judge is to produce norms by giving a meaning to normative utterances. In this sense, a judge is not only able to product normative meanings, but he has also the function of producing normative meanings. There are of course different conceptions of the activity of the judge such as that according to which a judge is *bouche de la loi*, or also that according to which the role of the judge is that to literally read laws. But I am convinced that these different interpretations of the judge's activity are not really convincing because of political reasons as well as because of language philosophy's arguments. First of all, in fact, the axiomatic thesis of Montesquieu according to which powers must be separate is not compatible with the thesis of a judge as *bouche de la loi*. Secondly, the theory of language is able to show that the thesis of the literally interpretation of laws is just a *flatus vocis*: the distinction between syntax and semantics, as well as that between denotation and connotation clearly shows that we can speak of a literally interpretation only if there is an unanimous consensus about the result of the interpretation. In this sense, we cannot have a literally interpretation of any legal system, even if this system is a very simple one such as the *Decalogue*. Also in this case, in fact, the problem of the derogation of God let some scholastics make a distinction between precepts, expressed in normative utterances, and norms as result of the interpretation (and in this sense we can remember either the position of Saint Thomas, either the opposite position of Duns Scotus).

But my ontological position is not only a realist one - as I have now explained. My position, in fact, is also that of a natural law philosopher. I do not only accept that norms exist as product of the interpretative action of judges, but I also believe that there exist natural laws, completely independent from the activity of the legislator. In fact, if we do not accept the existence of

⁹ A. Plantinga, *Warrant and the Proper Function*, Oxford 1993.

immutable norms, we cannot avoid that legal system becomes an arbitrary system. Positivistic philosophers, in fact, necessarily arrive at this consequence by denying the existence of immutable norms completely different from all the modifiable ones. But in this way, the ontology of a positivist position is not only a poor, but also a deficient one. In particular, my opinion is that we cannot arrive at justifying and founding the same universal rights if we do not accept the existence of immutable norms. According to some positivists, in fact, we can speak of universal rights only after their inscription into the Constitution. But this thesis is not able to give an account of our experience in that it can only reduced a fundamental text such as the *Déclaration* of 1789 into a sort of laic breviary. But the other paradoxical consequences is that the position of positivistic philosophers may seem not really different from that of Michel Villey¹⁰, even if Villey was surely a reactionary thinker. Also Villey, in fact, denied the fact that universals rights have an ontological reality, in that they are just subjective rights. In this sense the only difference between a positivistic position and a position *à la Villey* is that according to positivistic philosophers there isn't any immutable norm if it is not in a specific legal system, while according to Villey the only immutable norms are those linked to the commutative (thomistic) justice. So if the positivist ontology is a completely poor and deficient one, the ontology of Villey is not so poor, but it is never so rich to accept that there exist universal and fundamental rights. But I do not want now to exhaustively critic Villey: this is probably what I could do another time.

Now, I would only like to show that the positivistic position is not able to express our intuition according to which universal rights exist even if they are not expressed by constitutional norms. I agree in fact that universal rights are not protected in any legal system if they are not protected by constitutional norms. But I cannot agree that they exist only in that they are protected by constitutional norms. As well as I cannot agree with a particular formulation of the positivistic position such as that of Luigi Ferrajoli¹¹. According to Ferrajoli, in fact, a legislator is obliged to respect some rights by the same positive statutory law: a right of freedom, for instance, constrains not only a judge, but also a legislator when expressed in a constitutional utterance and then transformed in a constitutional norm. A judge is in fact obliged to protect its content; a legislator, on the other hand, is obliged to realise those conditions necessary to its applicability, even if the utterance of the right is absolutely general and not at all particular. But I cannot understand how it is possible to charge a legislator with this juridical duty (Ferrajoli says explicitly that it is a juridical duty, not a political one) in a positivistic picture: if this duty is written somewhere in the Constitution, it exists, but the theory of Ferrajoli is without object, because there is a new normative utterance, and not only a single utterance about a right of freedom. If this is not the case, I can speak of this duty as a political or a moral duty, but I cannot speak of it as a juridical duty: in a natural law picture, we can rightly say that this duty is juridical, but Ferrajoli wants to be a positivistic thinker and to state this duty as juridical. As an orthodox positivist can state, this is not a really positivistic sentiment: in my

¹⁰ M. Villey, *Droit et droits de l'homme*, Paris 1983.

¹¹ L. Ferrajoli, "Diritti fondamentali", *Teoria politica*, 14, 1998/2, p. 3-33.

opinion, this is a sign of all moral and political problems of a positivistic theory. In fact, it is impossible to avoid that legal system is an arbitrary system in a positivistic picture.

The real problem of any positivistic position is in fact its relation with the actual world. When we speak of universal rights we do not only speak of constitutional norms, because they exist even if and even when a positive legal system does not exist. But we all live in this actual world, where universal rights exist and have to exist. And for this reason I think that the positivistic conception of Law is an undesirable ideology.

On the contrary, the natural law position is not independent of our actual world. According to this position, in fact, immutable norms exist; they are super-constitutional norms; they can be thought as natural norms; they can finally be compared with the norms of the Christian natural law system, even if an important stream of natural law approach is today a laicist one, as Blandine Kriegel directly shows¹². But about this last point I do not want to say anything: the contrast between a thomist system and a system *à la* Scotus is something we cannot speak of now, also because not every natural law philosopher would accept a religious connotation of natural law.

REFERENCES

- Kriegel B., *Les Droits de l'homme et le droit naturel*, Paris 1986
 Collingwood R. G., *An Essay on Metaphysics*, Oxford 1940
 Dijon X., *Droit naturel*, I, Paris 1998
 Ferrajoli L., "Diritti fondamentali", *Teoria politica*, 14, 1998/2, p. 3-33
 Finnis J., *Aquinas*, Oxford 1998
 Kalinowski G., *L'impossible métaphysique*, Paris 1981
 Legendre P., *L'amour du censeur*, Paris 1974
 Legendre P., *Sur la question dogmatique en Occident*, Paris 1999
 Ollero-Tassara A., *Derechos humanos y metodología jurídica*, Madrid 1989
 Plantinga A., *Warrant and the Proper Function*, Oxford 1993
 Troper M., "Le norme sovraconstituzionali", *Analisi e diritto* 1996, pp. 255-274
 Villey M., *Droit et droits de l'homme*, Paris 1983
 Viola F., *Dalla natura ai diritti*, Roma 1995

Luca Parisoli

Centre de Théorie du Droit, Université Paris X

e-mail: luca.parisoli@u-paris10.fr

¹² B. Kriegel, *Les Droits de l'homme et le droit naturel*, Paris 1986.

THE NECESSITY OF IGNORANCE AND SOME PUZZLES CONCERNING IDENTITY AND INDIVIDUATION

ENRIQUE ROMERALES
Universidad Autónoma de Madrid

ABSTRACT: According to Timothy Williamson the best option to cope with the problem of vagueness is the epistemic theory. He introduces this theory as part of a general epistemic hypothesis, that there are some facts which are intrinsically opaque to our inspection. I think this hypothesis of the necessity of ignorance can be successfully applied to other topics in metaphysics; to classical dilemmas concerning personal identity and individuation, trans-world identity of merely possible individuals, trans-temporal identity and individuation of objects, and our knowledge of statements about the future. I take some of these cases to be examples of quasi non-removable ignorance, and most of them as paradigmatic examples of necessary ignorance.

KEYWORDS: Epistemic, ignorance, identity, individuation, knowledge, ontology, personal identity, possible worlds.

Since a decade ago or so Timothy Williamson has been defending what he has called the epistemic theory to cope with the problem of vagueness and related matters. According to this hypothesis, although there is a definite answer to all sorites-susceptible questions of the sort: “how many grains make a heap?” etc., there exists an intrinsic metaphysical limit that prevents us from ever to know what is the right answer to most questions of this kind. This is what he has called “the necessity of ignorance”. An hypothesis that, according to him, should be extended to other regions of philosophy¹.

Be it as it may for the specific case of vague concepts, I think the doctrine is applicable, and maybe even more plausible, to other metaphysical puzzles concerning identity and possible worlds. I will refer here to four issues of this kind, and eventually to a fifth of a different kind.

The first is a case involving important practical aspects: the problem of the legitimacy of abortion. Putting legal, social, psychological or theological

¹ His 1990, p. 103 ff.; 1992; 1994 chaps. 7-8; and 1996 develop this hypothesis.

issues apart, there is a decisive metaphysical problem that overrides all other matters: we don't know when a person begins to exist, among other reasons because we don't know with clarity and certainty what to be a person consists in. If, for instance, to be a person were to produce brain waves of such and such a kind, it would reduce to a medical problem: that of detecting when a foetus starts to produce such waves. But if to be a person is to have self-consciousness, the matter seems extremely difficult to solve by scientific means alone, since it is very difficult to determine in what exactly self-consciousness consists in, and also to check the exact time at which it appears. If, in the end, to be a person would not *consist in* anything, but is a purely metaphysical brute fact, such as to have an *haecceitas* or to be an individual essence *E*, then it would be intrinsically insoluble for science to determine when a foetus is a person (although it could there be some persuasive philosophical argument to locate the genesis of a person in some moment rather than in another).

You can have further complications if you take into account that, in addition to the metaphysical component, the concept of person has legal components, so that it might be proper to define it in modal terms, like this: *x* is a person if it is susceptible to develop such and such abilities (either brain waves, self-consciousness, moral awareness and responsibility or whatever). With this choice things would be far easier from a legal point of view, because independently of what to be ontologically a person consists in, a cygote is the thing which can become any of these, and, as a result, the cygote itself would be subject of moral rights. But this option is, no doubt, too restrictive, and offers a legal concept of person which is in conflict with other rights. And more importantly, the ontological issue remains intact.

In summary, as it seems not there to be any available means of determining what to be a person consists in, and as a result also no means of deciding when something begins to be a human being, the problem of abortion is radically insoluble, or so it seems. Some metaphysicians go a step ahead asserting that the concept of a person is merely conventional (for instance a legal one) and that there is no substantive ontological issue behind. Others assert that to be a person is not an all or nothing matter, but a gradual process. Both answers seem useless. The first, in addition to being very implausible, does not solve anything: for our conventional concepts—including the legal ones—are always grounded on objective facts (natural or social), even if their ultimate nature remains unknown or inaccessible to us. The second is yet worse: if persons are gradual entities, we continue with the problem of when exactly something begins to be a person even at the slightest level, and secondly, we have the problem of what percentage of personhood that something—supposedly a foetus, but maybe a baby—has to have in order for it to be liable to the rights proper for a person *per se*. So, this time the crucial problem would run as follows: how much of a person must have someone to be considered as a person?—unless you wanted to gradualize also

moral and legal aspects, and say that the more of a person you have, the more rights and duties; but this move surely would rend legal practices inviable.

It is here where the necessity of ignorance hypothesis offers an elegant and verosimile solution. In all probability, to be a person is an objective metaphysical fact. We have a rather accurate idea of what to be a person consists in (for example, Locke's definition in terms of a thinking and intelligent being, who possess reason and reflection and is capable of regarding itself as a same self at different places and times, to which moral components should be added), but we are ignorant of the ultimate metaphysical foundation of the fact of being a person. Maybe we can't even know whether there is such a foundation or the concept is as conventional as the concept of nation, to recall Hume's position.

Second case. What is at stake here is not so much what is to be a person, as what is it that makes that a person be precisely the one that he or she is, instead of any other one. As is pretty well known, personal identity raises a lot of puzzles, one is the case of fission. Example: if just after being fecundated the egg from which I in fact came out, my cygote, had split into two, A and B, which one would be me now? Parfit—as most philosophers—suppose that there are only four alternatives. I would be A; I would be B; I would be both; or I would be neither². Since, *ex hypothesi*, A and B are indiscernible, there is no more reason to say that I would have been A, than to say that I would have been B. Ergo, by the principle of sufficient reason, if I can not be one *rather than* the other, I can be neither A nor B separately. Obviously it is absurd to say that, in that case, I would now be both, A and B, for by the principle of the necessity of identity, one thing can not be (and also could not have been) two. It remains the fourth possibility only: that in that case I would not exist. This answer is usually admitted as the lesser bad. But it sounds extremely paradoxical to say that, if my cygote had split, I would not have existed. It seems much more plausible to claim that, in such a case, I would have had a twin brother. Notice that this fourth option conceals the highly implausible claim that for people who, as a matter of fact, don't have any twin brother, is metaphysically impossible to have had a twin brother. But it is surely true that there are possible worlds at which I both exist and have a twin brother. While the sentence "if I were you I wouldn't do that" seems senseless, because there is no coherent description of the state of affairs of me being identical with you, the sentence "if I would have had a twin brother, we would have had to share my toys" makes perfect sense, or so it seems.

The necessity of ignorance hypothesis give us again a clue. If my cygote had divided, as early as you wish in the process of my development, what would have happened is that *my* cygote would have been reduplicated, resulting in two *apparently* exact *replicas* of my original cygote. Now, does

² 1994, p. 256.

the question “which replica does bear numerical identity with the original?” make sense? It seems it does, because Parfit claims that the answer is “none of them”, although both are survivors of the original. And this answer is supported because we, supposedly, cannot conceive of any substantive difference between one and the other replica. But, what does ‘substantive difference’ mean here? It must mean either an ontological difference or a factual difference. And actually we can conceive of both, metaphysical and biological differences. In relation to the first, the original cygote might have an individual essence, an haecceitas, that, since haecceitas are non divisible, should attach only to one cygote after fission. Concerning biological differences, they are not only imaginable, but actual in most cases. True, the genetic code is the same in both cygotes, but the two sets of cells which make each new cygote just after fission are not exactly alike, to begin with, usually one has more cells than the other, and this difference could be relevant to be the bearer of personal identity with the former, even if we don’t know how. By saying that there is no relevant difference between the two because their genetic code is identical, you have already taken for granted that personal identity consist in the genetic code alone, but this is the point at issue. So, even if we are ignorant of any difference, this does not entail that there isn’t one, which could constitute the basis upon which personal identity holds and remains.

Third case. Let’s have a look to the existence of merely possible individuals and the supposed impossibility of referring to them³. Let’s imagine that Brad Brotherless has two sisters and no brother. But his parents got married very young, so they could have had more sons. In that case Brad would have had a brother –let’s call Bob the possible brother of Brad–. Unfortunately, the term ‘Bob’ is not a rigid designator, since by not existing Bob in the actual world, the name ‘Bob’ refers to different individuals in different possible worlds. Even in worlds where Brad has an only brother, that brother is in some worlds far younger than him, but in other worlds much older, in some much more intelligent, in others far more silly, etc. That means that Brad’s only brother is not the same individual in all the worlds at which this only brother exists. Now, let’s choose one world, for instance W78, in which Brad exists and has only one brother. Let’s call this brother ‘Bill’. Then the name ‘Bill’ would rigidly designate the Bob from W78, and so it could refer to Bill (Brad’s only brother at W78) across all possible worlds. Suppose further that in all worlds at which both Bill and Brad exist, Bill is four years older than Brad. Now Brad is able to say of himself, for instance, “if Bill had existed, I would have had someone who took care of me at school, someone

³ The existence of merely possible individuals is highly controversial. You can survey various arguments for their existence in Linski and Zalta, 1994, and Tomberlin and McGuinness, 1994.

who aided me with my homework”, and so on.

And here is the problem. We neither know nor are able to know who is Brad referring to by the name ‘Bill’, that is, with the utterance “Bob in W78”, because we do not know which world is W78. In fact, we are ignorant of the procedure according to which possible worlds have been ordered (although we have to allow that there is a sense in ordering them by some criteria or other). Granted the former, the necessity of ignorance has once more a lot of potential. As in other cases, from the fact that we are unable to find out, that there is an ignorance intrinsically unremovable, does not follow that there is no “fact of the matter”. There exists a possible individual –Bill– who inhabits some possible worlds, and that would have been a member of the actual world, should his parents (the same as Brad’s parents) have had sexual relations such and such a day and time –four years before they conceived Brad– and in such and such a way, in spite of it being intrinsically impossible for us to find out the precise moment and way in which, had anyone been engendered, this would had been Bill.

Up to this point the problems addressed have something to do with identity and individuation of people, real or possible. When we turn to identity and individuation of artifacts the necessity of ignorance hypothesis seems a bit less plausible. But I think that also here this hypothesis should be explored and exploited.

Let us focus on the issue of trans-temporal identity of objects, or its equivalent version in terms of possible worlds. I will change the standard example of the table for a computer. How many components of the computer can I replace (for another qualitatively equal) without the computer changing its numerical identity? Two claims seem beyond discussion: a) if I replace instantaneously all components, it is no longer the same, but another absolutely similar computer. B) If I replace no component at all, it continues to be the same computer, unless while all components work properly. But there are countless possibilities in between: from replacing just a key to replacing board, processor and hard drive. When does the computer cease to be the one it was and become a numerically different one? There is no clear answer to this question. According to mereological essentialism, what happens is that the computer, like any artifact, lacks *real individuation*, and for that reason does not have genuine identity, that is, metaphysical identity. Since it only has loose identity, for we attribute it in ordinary speech and thought, this sort of identity must have some metaphysically weaker foundation, and, lacking individual identity, it can only be some identity conferred upon the logical sum of all its parts at a given time. So, it is the identity of a set. But, as any set, its identity depends exclusively upon the numerical identity of each and every member. That means that, once a member is changed, the whole is also changed. As a result, for any artifact, if you change even a single component, the object is no longer metaphysically

the same.

If, taken as a doctrine to be applied in ordinary life, mereological essentialism seems extremely implausible, taken as a metaphysical doctrine makes perfect sense. Nevertheless, I think that the necessity of ignorance hypothesis again fares better. If one of the keys of my computer begins to fail 0.1 % times I press it, it sounds utterly implausible to say, even in metaphysical jargon, that my computer now is a numerically different object from the one it was before. Very likely the same holds if the only key out of order fails 0.2 % times, and so on. To be sure, the day at which every key fails always, my computer is no longer the same –to begin with, because it no longer is a computer. Now, maybe there is a point in between at which my computer ceases to be the one it was and becomes a different one, although we are ignorant of where that point lies. At this point someone might protest: in the first case we didn't know when does exist a person because we are ignorant of what to be a person consist in, but could we say in the same vein that we are ignorant of what to be a computer consist in, since it is us who have invented it? This question has two answers. On the one hand, we know perfectly well what to be a computer consist in, what components it must possess, how they are to work, etc. But when metaphysical issues arise we are not faced with a technological problem, but with a conceptual problem. We actually don't know what to be a computer consists in, in the sense that we have not made up a concept so complete and accurate that it allows us to settle beforehand any possible puzzle about any conceivable counterfactual situation a computer can be in. And this happens with almost any concept. For that reason almost all concepts are more or less sorites susceptible, even the ones which are not properly vague. Note that it is no wonder that we are capable of make a computer without possessing a clear concept of what to be a computer consist in. We also know pretty well how to make a person in spite of lacking any clear idea about what its identity consists in.

Now, for all this array of questions concerning identity and individuation a very different kind of answer is available to us: that identity in some cases is intrinsically indeterminate, or to speak in Terence Parsons terms, that there are entities without identity⁴. This claim is particularly appealing when we are talking of artifacts, and even more with cumulative entities like heaps or fuzzes. No doubt, heaps and fuzzes do exist, indeed lots of them exist. But they don't seem to have any real principle of individuation, so maybe they are indeed, *pace* Quine, entities without identity. But, once you have made room for some entities without identity, why not spread the scope of this ontological category beyond cumulative objects and say that many more entities that we thought, maybe even all entities, are in the last analysis

⁴ Parsons, 1987. A full developed version of this theory appears in his 1980, which he sees as a rehabilitation of Meinong's views.

like this?

I think this approach is suggestive and deserves closer scrutiny. But in the end it involves so a radical upsetting of basic notions from classical logic and from our best established ways of inference that, as Williamson has repeatedly insisted, you have to pay too much. To allow for entities without identity is to allow for objects intrinsically vague, and I think this is eventually incoherent. To my mind, the best procedure is to say that, although these sort of entities undoubtedly exist in the loose and popular sense of 'exist', they do not exist in the strict and philosophical sense of 'exist', which should be applied only to real individuals (and maybe to properties and relations too). For that reason I understand that the necessity of ignorance hypothesis, in having a far lesser epistemic cost, is a better explanation and should be preferred here. We are entitled to say: I don't know when my computer ceases to be the same it was, because I don't have a precise and detailed enough concept of computerhood, but there must be a point beyond which it is no longer the same. To be sure, when it is definitively broken, it is no longer the same, because it is no longer a computer, except in appearance. But the ignorance in this case seems different from the former cases. Here we have a lack of conceptual discrimination that could be removed by further investigation if we wished (and I think most cases of vagueness are also like this). But in the first three cases we had something more, we had ontological aspects which seemed beyond our scope, outside our cognitive powers. So I would speak here of *contingent ignorance* rather than of necessary ignorance.

Finally, I would like to add a fifth and last case, having nothing to do with identity and individuation, and for that reason maybe less controversial. Think of statements concerning the future and let's suppose that the future is not totally determined, which is quite a bit plausible after all. Now, let's take some contingent statement about the future, for instance, Barcelona FC will win Champions League of year 2.000. Either this statement has a truth value or not. If it does have a truth value, we are ignorant of it, and for us it is ontologically impossible to know this truth value here and now. An omniscient being would know. If, on the other hand, it lacks any truth value, we are not properly ignorant of it, but nevertheless we continue without knowing whether Barcelona will be champion or not. Only that in this case we are not alone in our ignorance: even an omniscient being does not know the answer to "who will win Champions League of 2.000". But the crucial point is that in either case there is an ignorance we ontologically cannot remove; at least not here and now.

Nevertheless, concerning this case someone might reply like this: if contingent propositions about the future do lack truth value, then we are not properly ignorant of anything when we are unaware of their truth value. So, at best we don't know whether we have here a necessary or non-removable ignorance, or just an appearance of ignorance. And maybe we will eventually

be capable of find out that there was nothing to be known. I think this line of argument is wrong. If contingent propositions about the future do lack truth value, the situation is not that we are less ignorant than supposed. We continue without knowing, and without being able to know by any means whatsoever (except by waiting the required time) the right answer to “who will win the championship?” and related matters. But questions about the future are proper quite independently of whether they can or cannot be answered in principle—compare “are there facts unnoticeable to everybody?”, which seems to be perfectly proper—. The only important thing this time is that we are not alone in our ignorance: an omniscient being is in a far better position to bet, but in an equally bad position to properly know. So, I would claim that, if there is some uncontroversial example of non-removable ignorance this is our ignorance about what will happen in the future. It is intrinsically impossible for us to know today what will be Dow Jones index tomorrow, but there is a correct answer to that question.

Let me finish with some remarks concerning the relation between truth and verification. The days in which meaning was defined in terms of verification are gone away, but it is not unusual to define truth in terms of some conception of verification or other. By saying that *P* is true if and only if it is verifiable in principle, you can mean many different things. One is that if *P* is true an omniscient being must be in a position to know that *P*, and conversely, that if *P* is not true an omniscient being must be in a position to know that not *P*. This is rather plausible, if only because an omniscient being is an ideal knower who, as such, must know the truth-value of every proposition (and also the lack of truth value of any proposition, in case there are any). But to define truth in terms of what an omniscient knower does know (or just in terms of what he *can* know, if he is also omnipotent and can refrain from knowing, if he wish) is perhaps good from a logical point of view, but of no practical consequence, for, after all, who knows what an omniscient being can know?

But by ‘verifiable in principle’ you can mean that we, human beings, are or will be in a position to know, in case some conditions are met. And the point is that, concerning the former examples, we are not in a position to know (and concerning the first three we also will never be in a position to know). Sometimes due to ontological grounds, like in the case in which you have to cope with the individuation of a *haecceitas* or with the knowledge of the truth value of a contingent statement about the future, sometimes due to technical but non-removable difficulties, like in the case of the split of a cygote, and sometimes due to one of the two (or maybe to both), and you cannot know which, like in the abortion case.

So, although there is a strong connection between meaning and truth, the link between truth and verification is logical, not practical. If some proposition is true, then there is something to be know, but this does not mean

that there must also be someone who is actually capable to discover that truth.

In the cases involving identity and individuation one solution was to regard some entities as lacking actual individuation and, consequently, as lacking identity in the "strict and philosophical sense" (to use Samuel Butler's own words, now recalled by Chisholm⁵). In this case, concerning statements about the future, we have as a solution to take them as indeterminate, and so as lacking truth value. Both claims are highly problematic. Both bring about a full revision of classical logic, both imply rejecting some of our better established patterns of reasoning. The necessity of ignorance thesis avoids all of this difficulties, and appeal always to the same general principle: there are some ontological facts we are not in a position to know. Further, for some of them, we will never be in a position to know. With many extremely important matters we are, to borrow Williamson phrase, epistemically homeless. We have to take epistemic risk. This offers a more plausible explanation of all the mentioned puzzles. If we take also into account that this hypothesis applies successfully to a wide variety of cases, I think we have additional grounds to continue exploring this fruitful theory.

REFERENCES

- Care, Norman S. & Grimm, Robert H (eds.): *Perception and Personal Identity*. Cleveland, The Press of Western Reserve University, 1969.
- Chisholm, Roderick: "The Loose and Popular and the Strict and Philosophical Senses of Identity". In Care & Grimm, 1969.
- Chisholm, Roderick: "Identity Through Time". In Keifer & Munitz, 1970.
- Chisholm, Roderick: *Person and Object*. La Salle, Open Court, 1976.
- Kiefer, Howard & Munitz, Milton (eds.): *Language, Belief and Metaphysics*. Albany, State Univ. of New York Press, 1970.
- Linsky, Bernard & Zalta, Edward: "In Defense of the Simplest Quantified Modal Logic". In Tomberlin, 1994; 431-58.
- Parfit, Derek: *Reasons and Persons*. Oxford, Clarendon, 1984.
- Parsons, Terence: "Entities Without Identity". In Tomberlin, 1987; 1-20.
- Parsons, Terence: *Nonexistent Objects*. New Haven, Yale University Press, 1980.
- Tomberlin, James E. (ed): *Philosophical Perspectives*, 1. Atascadero, CA, Ridgeview, 1987.
- Tomberlin, James E. (ed.): *Philosophical Perspectives*, 8. *Logic and Language*. Atascadero, CA, Ridgeview, 1994.
- Tomberlin, James E. & McGuinness, Frank: "Troubles with Actualism". In Tomberlin, 1994; 459-66.
- Williamson, Timothy: *Identity and Discrimination*. Oxford, Blackwell, 1990.
- Williamson, Timothy: "Inexact Knowledge". *Mind*, CI, 1992; 217-42.
- Williamson, Timothy: *Vagueness*. London, Routledge, 1994.

⁵ Chisholm has insisted on the necessity of distinguish between these two senses of identity in many different places. Among them in his 1969 and 1970. A fully developed version of the fundamental distinction between the kinds of identity and individuation of objects and people appears in his 1976. The idea is in the end Leibnizian: that only minds or souls are real individuals. The expressions come from Joseph Butler's "Of Personal Identity", included in his *The Analogy of Religion* (London, 1736).

Williamson, Timothy: "Cognitive Homelessness". *The Journal of Philosophy*, 96, 1996; 554-73.

Williamson, Timothy: "La necesidad de la ignorancia. Lógica, indeterminación y metafísica". *Cuaderno Gris*, 4. Madrid, UAM, 1999.

Enrique Romerales

Universidad Autónoma de Madrid

e-mail: enrique.romerales@uam.es



TWO PROPOSALS FOR FORMALIZING A BUNDLE THEORY BASED ON UNIVERSALS

CHRISTINA SCHNEIDER
Universität München

ABSTRACT: The paper presents two proposals for a formal bundle theory. One proposal is based on a "compresence relation". The second proposal is more sophisticated: it formulates the "bundling together of universals" by topological means. The second approach, however, gives a basis of understanding space-time events as being "constituted out of universals". Finally, it will be argued that none of the two proposals, taken in isolation, seems to give rise to an interesting ontology.

KEYWORDS: Bundle theory, immanent universals, platonic universals, individual(s), topology.

Recently, the metaphysical theory that universals are the primary constituents of "the world" and that "individuals" are "bundles of universals" regained attention. This was due to papers by J. O'Leary-Hawthorne and J. A. Cover (O'Leary-Hawthorne [1995], O'Leary-Hawthorne and Cover [1998])). O'Leary-Hawthorne and Cover call their concept of universals "immanent universals" to distinguish it from what they call "platonic universals". Their universals, in contrast to platonic ones, inhabit "the world"; platonic universals according to those authors remain outside "the world".

The critics of this conception, e.g. MacDonald [1998], Zimmerman [1997] and Vallicella [1995] focus their doubts mainly on the claim that this conception is committed to the Leibniz Principle, i.e., to the "Principle of the Identity of Indiscernibles". "Black's world" counts as a perspicuous support of that claim. It must be admitted, however, that Cover's and O'Leary-Hawthorne's reply to this argument is not convincing.

But, before any argument concerning this topic can be considered meaningful, a clear concept of how to conceive of "bundles of universals" and "bundles of universals which constitute individuals" should be available.

This paper attempts to present two formal concepts for "bundling universals" and "individuals as bundles of universals": one for "platonic

universals" and one for "immanent universals". As a result, it can be seen that the Leibniz Principle does hold naturally for individuals as bundles of immanent universals, whereas individuals as "bundles of platonic universals" are independent from Leibniz Principle.

The article is organized as follows: The first section presents a formulation of platonic universals, bundles thereof and their relation to the region where they are "instantiated". The second section gives a formulation of immanent universals and of individuals as bundles thereof. The third and last section sketches the thesis that neither a pure theory of platonic universals nor a pure theory of immanent universals gives hope for a satisfying ontology. If basing an ontology on "universals" alone, there should be involved both aspects.

Both concepts are formulated only for instantaneous aspects/individuals. The task of formulating "change" is not pursued in this paper, since none of both concepts alone seems to come to grips with "change".

1 Platonic Universals

Platonic universals presuppose, at least implicitly, a dimension "beyond the universals": the dimension where these universals, single or "as bundles", are "instantiated". Often, this further dimension is conceived of as "space" or "space-time" which furnishes the "places" where universals or bundles thereof are "instantiated". In what follows, however, no such strong assumptions are made. As will be discussed in the last section, they are problematic in a frame which claims that universals are the primary constituents of "the world" or of "reality". Therefore, in what follows, this "instantiating" dimension is simply a set, W , and the elements of this set can be seen as "places" (not to be read as spacial places or space-time-events) where universals or bundles thereof are "instantiated".

The other dimension is the "platonic heaven" where the universals and bundles thereof, conceived of only as bundles of universals and not yet as individuals, are situated. This region of universals must be endowed with a structure which gives rise to a non-arbitrary bundeling of (simple) universals. For short, a "platonic heaven" is a set of simple universals endowed with what will be called a bundling structure. This structure will be explicated subsequently. Afterwards, the individual instantiation function (instantiation function(s)) and the exemplification relation will be defined. This leads to the definition of individuals as instantiated bundles of universals.

In defining the bundling structure an informal intuition of how to frame bundles of universals which are "individual-inductive" is at work. This intuition is twofold:

1. "Individual-inductive" bundles should be "maximal" and "complete" in the following sense: there can be added no further universal without

running the construct into "inconsistencies" and, conversely, leaving out one (or more) universals from the construct such that no entity which may be called "individual-inductive" is left. What is left may be seen as a complex universal. This intuition is also basic for the compresence relation $\text{citeN}\{\text{Russelluniv}\}$ introduces.

2. Individual inductive bundles may share universals.

Putting these intuitions together leads to the subsequent definition of a bundling structure and individual-inductive-bundles.

Definition 1. Let U be a set of simple (platonic) universals and let $\mathbf{I} \subseteq \wp(U)$ be a family of sets for which the following holds:

1. $\bigcup_{J \in \mathbf{I}} J = U$
2. For $u \in U$ define $\mathbf{I}_u := \{J \in \mathbf{I} : u \in J\}$ and denote the elements of $\mathbf{I}_u$ by I_u ; then, for $J \in \wp(U)$:

$$J \in \mathbf{I} \Leftrightarrow J = \bigcap_{u \in J} \bigcup_{I_u \in \mathbf{I}_u} I_u$$

$(U, \mathbf{I})$ might be called bundling structure, the elements of $\mathbf{I}$ are called individual inductive bundles and for each universal $u \in U$ the set $b(u) := \bigcup_{I_u \in \mathbf{I}_u} I_u$ is called bundling neighborhood of u . This set can be regarded as the set of all universals which can be bundled with u .

To see that the maximality intuition is met, notice that for $J \in \mathbf{I}$ and $v \notin J$ the set $J \cup \{v\}$ is not an element of $\mathbf{I}$, otherwise $J \cup \{v\} = J \cup_{I_v \in \mathbf{I}_v} I_v$ must hold. This contradicts the assumption.

On the other hand, let be $v \in J$ and $J \in \mathbf{I}$, then, if $J \setminus \{v\}$ would be an element of $\mathbf{I}$, then $J \setminus \{v\} = \bigcap_{u \in J \setminus \{v\}} \bigcup_{I_u \in \mathbf{I}_u} I_u \supset J$ holds. This a contradiction.

To see that the third intuition can be met, regard the following "mini-model": Let be $U = \{a, b, c\}$ and $\mathbf{I} = \{\{a, b\}, \{b, c\}\}$, then $\mathbf{I}_a = \{\{a, b\}\}$, $\mathbf{I}_b = \{\{a, b\}, \{b, c\}\}$ and $\mathbf{I}_c = \{\{b, c\}\}$ and $b(b) = \{a, b\} \cup \{b, c\} = \{a, b, c\}$ and $\{a, b, c\} \notin \mathbf{I}$.

1.1 Individuals and instantiated universals

Let W be the region of "instantiating places" and suppose that this region is not endowed with any structure, then a function:

$$ii: W \rightarrow \mathbf{I}$$

will be called *individual-inductive-function* and each pair $(w, I) \in W \times \mathbf{I}$ with $ii(w) = I$ will be called an individual. Since ii is a function, it is perfectly possible that for different w and w' $ii(w) = ii(w')$, therefore the Leibniz Principle does not hold generally in this frame. If one is committed to the Leibniz Principle, one can, of course, define ii to be one-to-one.

The definition above is rather simple, it reflects the assumption that there are no places that do not instantiate an individual. If one does not share this assumption, ii should be defined as a *partial* function. The definition does not presuppose a determinate solution to the question whether all individual-

inductive-bundles are instantiated or not. If one is committed to the claim that any individual-instantiating-bundle is instantiated, *ii* should be an onto-function.

So far, no notion has been presented which formulates the "instantiation" of universals, single or complex, which are not individual inductive. Any such notion depends on presuppositions which are listed below, without claim of exhaustiveness:

1. Universals are only instantiated as coconstituents of individuals.
2. It is possible that universals are instantiated without coconstituting individuals; they can be "free-floating".
3. Any set of universals may be instantiated at one place.
4. There are restrictions which regulate which universals may be instantiated at one place.
5. Any place in W instantiates some universals.
6. There are "universal-empty" places in W .

The first claim leads within the frame presented so far to the following definition:

Definition 2. Let i_λ be a family of functions from W in $\wp(U)$:

$$i_\lambda : W \rightarrow \wp(U)$$

with $i_\lambda(w) \subset ii(w)$ for each λ , then $i_\lambda(w)$ is an universal instantiated at w . If $i_\lambda(w)$ is a singleton, then the instantiated universal is a simple universal.

If the intuition formulated in clause 6 above should be met, then the ii_λ are partial functions and their domain(s) are the domain of the corresponding partial function *ii*.

If the second intuition is shared, there remain the options whether any combination of universals can be coinstantiated or whether there are restrictions concerning coinstantiatedness. The trivial option that there are no restrictions at all leads to a family of functions $ii_\lambda^t : W \rightarrow \wp(U)$.

The fourth intuition leads to two options, comparable to the question whether there should be only "conjunctive universals" or also "disjunctive universals" (compare, for instance, Amstrong [1978], p. 19-p. 23). Without going into details, it should be mentioned that both options can easily be captured by modifying the instantiation-functions with respect to restrictions concerning their range.

For each (family of) instantiation-functions, each expressing one or the other of the intuitions listed above, one can define the *exemplification-relation* simply as the inverse image of the respective functions.

2 Immanent Universals

The main intuitions concerning *immanent universals* is that they are "in" the world and that there are only universals as the primary constituents of the world. Putting these two intuitions together leads to the thesis: the world is

nothing over and above universals and complexions thereof. From this it follows that there is no "region of the world" which is not a complexion of universals. Admitting of "universal-empty regions" within that frame would amount to postulate some entity or entities which are neither universals nor complexions thereof; this contradicts the mono-categorical intuition governing a bundle theory.

In analogy to platonic universals, there are some intuitions of how universals relate to each other and of how individual entities, individuals and what could be called "places" are to be conceived of:

1. One universal can coconstitute several individuals.
2. Individuals are maximal and complete.
3. Not all universals which are "together" make up an individual.

In contrast to platonic universals, however, any "place" which is constituted out of immanent universals is the "place" of an individual. This will be problematized in the last section.

Before starting to give a formal account of universals, individuals and "places", some more precise intuitions concerning how universals may "combine" and how to conceive of individuals are given.

2.1 The Space of Immanent Universals

1. Finite conjunctions of universals are universals; non-finite conjunctions of universals may be individual entities (but no "robust individuals").

2. If some (finite) "conjunction" of universals should be mutually "incompatible", this "construct" should be conceived of as a limiting case, an "empty universal". Moreover, there should always be some "incompatible" universals; the "world" should be rich enough that not anything can go together with anything else.

3. (All) adjunctions of universals are universals. But not every adjunction, especially no disjunction, can --- as a whole --- coconstitute an individual.¹

To make these intuitions precise and to get a strong tool at hand, one can use set-theoretic language; the intuitions formulated above now read as follows:

Definition 3. Let U be a set of sets, the set of universals, for which the following holds:

¹ This is in accordance and partly in contrast with the claim of Armstrong [1978], p. 18ff that disjunctive universals cannot, as disjunction, be universals "instantiated in" an individual. The claim above respects this. But it claims further that this impossibility does not prevent a universal from being a universal. Armstrong, however, locates his conception of universals within a two-categorical frame which is crucially different from "bundle theories".

1.

$$U_1, U_2 \in \mathbf{U} \Rightarrow U_1 \cap U_2 \in \mathbf{U}$$

2.

$$\exists U_1, U_2 \in \mathbf{U} \quad U_1 \cap U_2 \in \emptyset \quad \emptyset \in \mathbf{U}$$

3. For any family (U_i) in $\mathbf{U}$

$$\bigcup U_i \in \mathbf{U}$$

Therefore, $W := \bigcup \mathbf{U} \in \mathbf{U}$

The pair $(W, \mathbf{U})$ is called space of immanent universals.

2.2 Individuals as Bundles of Immanent Universals

With the space of immanent universals, the frame for constituting individuals is given. Besides the intuition of maximality and being a bundle of universals, two other intuitions about individuals should be met: individuals cannot "overlap" as a whole, i.e., they cannot share "the same place" and individuals are in some sense bound to places. These intuitions are met by the next definition:

Definition 4. Individuals are subsets $I \subseteq \mathbf{U}$ for which the following clauses hold:

1. Individuals are made up by compatible universals:

$$I \neq \emptyset$$

$\cap I$ is called place of the individual I .

2. Maximality:

$$\forall U \in \mathbf{U} \quad U \cap \cap I \neq \emptyset \Rightarrow U \in I$$

3. Individuals cannot "overlap": For individuals $I_1, I_2 \in \mathbf{U}$

$$I_1 \neq I_2 \Rightarrow \cap I_1 \cap \cap I_2 = \emptyset$$

For each individual I the set $\cap I$ is called place (of the individual I).

4. There is no "individual free region" in the "world":

$$\bigcup_I (\cap I) = W$$

The restriction on individuals formulated in clause 4 seems to be the most problematic one². If one would restrain from it, one would still get a consistent structure. This structure, however, would have some serious drawbacks, as will be sketched in the next section.

Some consequences of the definitions above should be drawn:

1. Given any individual I , for any $U \in I$ $U \neq \emptyset$ holds. This means that inconsistencies cannot coconstitute individuals.

² Whether it can be fulfilled depends on the space of universals $(W, \mathbf{U})$. Therefore the restriction ought to be defined when defining the universal space. But, since it becomes clearer in connection with individuals, it was given in that place.

2. Given any individual **I** and universals, U_1, U_2 coconstituting it, then $U_1 \cap U_2 \in \mathbf{U}$. Since $U_1 \cap U_2$ is an universal, this may be judged as superfluous. But notice that those individuals which are coconstituted by U_1 or U_2 , but not of both, are not coconstituted by $U_1 \cap U_2$. Therefore, this "complex" universal contributes to individuals in a way which is not reducible to the contribution of U_1 and U_2 .

3. Given any individual **I** and any family U_i of coconstituting universals, then $\cup U_i$ is an universal coconstituting **I**. Note that no such adjunction is a disjunction. As adjunctions, in coconstituting individuals different from their adjuncts, they behave in the same way as was indicated in the previous clause.

4. The Leibniz Principle does hold for individuals, since two individuals are identical if and only if they are coconstituted by the same universals.

5. "Places" where the individuals are located are determined by the respective individuals and not vice versa.

6. There are no "places" in the world devoid of any individuals.

It should be noted that albeit set-theoretic language was used in formulating the space of immanent universals, no use was made of "elements of sets". One could have defined the structure by "algebraic " means too. Sets are, however, adequate to deal with space-time-structures.

3 Some Final Considerations

Beyond any doubt, the formulation of *platonic universals* gives more means at hand to compete with several intuitions about a "world made up by universals". On the other side, it makes use of some problematic structural features: a) The "platonic heaven" is constituted in a combinatorial manner, based on *primitive* (simple, not analysable) universals. A feature like this was problematized by ?], p. ...ff. b) The set W , the "places" where universals are instantiated, is neutral with respect to the universals instantiated by them. Feigning for a moment that "the world" is made up by individuals only, any permutation of W would lead to the same "world". The elements of W are simple and bare individuators. c) In this respect they remind of "haecceities", "bare particulars" or "bare substrata", inheriting all the intelligibility problems of these entities. In that respect, they undermine tacitly the mono-categorical attempt of a bundle theory.

Immanent Universals, on the other hand, are committed to two rather strong principles: the Leibniz Principle and the claim that there is no region devoid of individuals in "the world". Skipping clause 4 of the definition of individuals would leave some complexion of universals, not leading to something that could be called "place".

There is a problem, however, which seems to plague both conceptions: the problem of how to conceive of space and time, or space-time. Is the geometry of space-time constituted by platonic universals? If it is, it seems not to fit the combinatorial schema: starting with simple universals and bundeling structures of higher complexity. Should the geometry of space-time be conceived of as an axiomatic system and the W , endowed with a further structure, as a representation thereof? This structure, however, can hardly be seen otherwise than as "immanent". Further, it is not at all clear whether a line, for instance, as structure of W is an individual or an universal entity.

Concerning immanent universals, the situation seems to be inverse: It may well be possible to enrich the structure of the space of immanent universals by geometric structures. But, what would be the status of an axiomatic formulation thereof? The entities this formulation deals with can hardly be conceived of as "immanent". What ontological status can be attributed to them? Are they mere mind-dependent "abstractions"? Therefore, regarding an enriched space of immanent universals as a representation is problematic.

Without being able to give a detailed account of it in this paper (see Schneider [1998], Schneider[1999]), it should be remembered that there is a model of how to conceive of space-time in an ontological frame based on universals: simple models of General Relativity Theory. In this frame, the space of immanent universals, with some minor modifications, figures as a *differentiable manifold*. The refined geometry can be conceived of as an axiomatic system and adequate representations thereof can be conceived of as the *tangent space(s)* of the manifold. In this respect, they can be seen as "universal". The link between the tangent space(s) and the manifold is formulated by the *diffeomorphisms*. In this respect the space of immanent universals is not neutral with respect to the tangent space(s). (Compare, for instance, Sachs and Wu [1977] p. 254 f.) Individual features of the geometry can be seen as values of vector-/tensor-fields at specific points or places of the manifold: the space of immanent universals.

In the light of the sketchy remarks above, it can be judged, however, that neither purely immanent universals nor purely platonic universals seem to be able to give any interesting ontology.

Further, no combinatorial ontological architecture seems to give hope for a refined ontology: Those ontologies are based on fully determined entities (simple universals, for instance), which are conceived of as intelligible and meaningful on their own and in isolation, and which are "combined" and "bundled" (under some pretheoretic paradigm of collecting and combining) to complex entities. There is, e.g., an euclidean or lobachevskian line, dependent on the entire geometry in question, but there is no line *tout court*.

Moreover, the distinction between "immanent" and "platonic" universals seems not to be helpful in theoretic respect: for instance, what status should be attributed to diffeomorphisms "joining two spheres"?

Positively, an ontology based on structures which can be regarded as being "universal" and in whose frame individual entities are derivative substructures seems to be a feasible and interesting project.

REFERENCES

- Armstrong, D. M. [1978]: *A Theory of Universals*. Cambridge: Cambridge University Press.
- Macdonald, C. [1998]: Tropes and Other Things, , pp. 329-350.
- O'Leary-Hawthorne, J. [1995]: The bundle theory of substance and the identity of indiscernibles. *Analysis* 55, pp. 191-196.
- O'Leary-Hawthorne, J. and Cover, J. A. [1998]: A World of Universals. *Philosophical Studies* 91, pp. 205-219.
- Russell, B. [1911]: "On the Relation of Universals and Particulars". *Proceedings of the Aristotelian Society* 12, pp. 1-24.
- Sachs, R. K. and Wu, H. [1977]: *General Relativity for Mathematicians*. New York: Springer-Verlag.
- Schneider, C. [1998]: "Leibniz's Theory of Space-Time: An Approach from His Metaphysics. *The Monist* 81, pp. 612-632.
- Schneider, C. [1999]: *Ein formaler Zugang zu Leibniz' Metaphysik*. Unpublished manuscript.
- Vallicella, F. W. [1995]: Bundles and indiscernables: a reply to O'Leary-Hawthorne. *Analysis* 57, pp. 91-94.
- Zimmerman, D. W. [1997]: Distinct Indiscernables and the Bundle Theory. *Mind* 106, pp. 305-309.

Dr. Christina Schneider
Universitat München
e-mail: schneider@statistik.uni-muenchen.de



VERTICAL DEPENDENCIES AND THE EXCLUSION PROBLEM

AGUSTÍN VICENTE

Universidad del País Vasco/Euskal Herriko Unibersitatea

ABSTRACT: The by now famous exclusion problem for mental causation admits only one possible solution, as far as I can see, namely: that mental and physical properties are linked by a vertical relation. In this paper, starting from what I take to be sensible premises about properties, I will be visiting some general relations between them, in order to see whether, first, it is true that some vertical relation, other than identity, makes different causations compatible and second, whether physical and mental properties can be pairs of such relation.

KEYWORDS: Problem of causal exclusion, properties, relations between properties, supervenience, realization, genus-species, determination.

1. Introduction

In this paper I will be dealing with the well-known problem of exclusion for mental causation. Such problem, said to have been first introduced in contemporary philosophy by Norman Malcolm, is usually presented (following Jaegwon Kim) as a conflation of three well-entrenched ideas, namely

(i) principle of the causal closure of the physical: every physical effect has a causal antecedent that is also physical;

(ii) causal efficacy of the mental: mental events produce changes in the physical world;

(iii) principle of the causal-explanatory exclusion (PEE): there cannot be two causes/causal explanations that are both complete and independent for one event, except in cases of overdetermination.

In what follows I will assume that all three ideas are really well entrenched, so that the denial of any one of them would make some (big) distortion to our preferred view of the world. I will further make some other assumptions, some of them plausible as they stand and some others a bit more

controversial. First, I will understand that events are complexes of particulars, properties and times, that is, an event is the instantiation of a property by a particular at a time. Properties, on the other hand, will be understood as universals, though I hope this one is not a substantial assumption, i.e. one that makes a big difference to the outcome. Further, I will take properties to be entities with an internal structure, and detachable from their causal powers, that is, properties, on my account, are not causal powers, and do not have such powers either essentially or necessarily. This is a highly controversial assumption and besides one that could happen to be substantial in the sense just noted. The only reason to take it as a standpoint is that the discussion on the exclusion problem does normally take it for granted (Kim does, at least), and what I want to do here is see how far we can go starting from some plausible and shared ideas to the solution of the puzzle. Thus, I would have no problem -or so I think- in removing this pillar if it happens to be the one causing trouble. For the time being, however, let's think of properties as, though epistemically picked up by the causal powers they endow particulars with, not individuated by such causal powers.

More assumptions and/or caveats: The problem of causal exclusion consists in the difficulty of making it compatible that both mental events and physical events cause some given effects: the same effects. That is, I exclude from the start the possibility that mental and physical causes act at different levels, bringing about different kinds of effects. Attending to this "dual-explanandum" strategy would take us too far away, but I can confess that, even though I see it with some sympathy, I am not sure I entirely understand it.

As said above, I want to accept all three propositions, so I will not pay attention, at this stage, to those that start by questioning one or another. This means that the possible solutions to the problem are heavily restricted: as a matter of fact, they are restricted to only one. This "one and only" possible solution will be called here 'the vertical relation approach': First, as the principle of causal-explanatory exclusion allows us to have two conflicting causes-explanations only if we discover that one completes the other, or that one depends on the other (or, third, if there is overdetermination, but there is not), and by hypothesis, physical causes are complete, the problem of mental-physical exclusion can just be solved by means of a relation that establishes a dependency between the two causal stories, mental and physical. Second, this relation can only be "vertical". Otherwise, it would imply the rejection of our principle (i) above, for it would mean either that the physical cause was caused by the instantiation of a mental property, or that the instantiation of

this mental property, although caused by the instantiation of a physical property, was the immediate antecedent of the physical effect. The first situation straightaway implies the rejection of the principle of the causal closure of the physical world, as stated above. The second, on the other hand, implies the rejection of a perhaps stronger principle that we would also subscribe, namely, that all physical effects have an immediate causal antecedent that is also physical.

This brief discussion of the vertical relation approach brings us to the end of this introductory section, that I will close with a final caveat: I will assume, along with many others, that mind-body multiple realizability, or, to put it in neutral words, the fact that particulars with very different physical properties can nevertheless instantiate the same mental properties, excludes the identity relation as a candidate to our vertical relation approach. I assume that such fact is indeed a fact and that it counts decisively against identity theories. Of course, this last assumption is as substantial as one could possibly imagine and still very controversial. But it is part of the game I will be playing, the most plausible game in town, I hope.

2. Failed attempts: supervenience and realization

I will start by studying the prospects of two relations between properties that have been proposed in one moment or another, as solutions -of the type just explained- to the problem of causal exclusion. Such relations are supervenience and realization. My conclusion will be pessimistic: supervenience, as has been remarked by Kim and others, cannot in fact be rendered as a “deep” relation and does not, by itself, tell us anything concerning exclusion problems, whereas realization, though specific and deep enough, cannot provide realized properties with causal powers. As what I am about to say about both relations is not original, I will try to be brief in the discussion; realization, however, is more intriguing than supervenience, so its discussion will take a bit longer.

2.1. Supervenience

The usual concept of supervenience involved in discussions of mental causation is the variety strong supervenience. Its well-known definition is the following:

(1) One level of properties *strongly supervenes* on another if and only if, necessarily, for every x and every higher level property M , if x has M , then there is a property F of a lower level, such that x has F , and necessarily if any y has F , then it has M .

For about a decade, the problem of exclusion was thought to be solved by means of this dependency relation: if mental properties strongly supervene on physical properties, it was said, then they can be causally efficacious, because they have been shown to be dependent on physical causes.

This model, first proposed by Kim (1984), had some problems, some internal and some external. One external problem is the so-called problem of externalism: as a matter of fact, this problem says, mental properties do not supervene on the physical properties whose instantiation is causally responsible for the instantiation of behaviors; that is, the antecedent in the conditional above is false and hence the model can't be applied.

One internal problem, on the other hand, is that supervenience cannot be a sufficient condition for causal efficacy; otherwise all the properties that supervene on a certain neurological property would have to be regarded as causes of the behavior that is an effect of the instantiation of that property.

However, hard as these problems may look, there is one, or better, a family of them, that are even harder. Kim (1993; 167) complains about these problems in the following way:

"[S]upervenience itself is not an explanatory relation. It is not a "deep" metaphysical relation; rather, it is a "surface" relation that reports a pattern of property covariation, suggesting the presence of an interesting dependency relation that might explain it [...] It is a "phenomenological" claim, not a theoretical explanation."

We can say that the "hard problem" for the supervenience account of mental causation lies in the fact that the definition stated above is too coarse-grained. It is compatible with a set of situations that do not solve the problem of mental causation, such as emergentism and type-epiphenomenalism. That is, it does not discern the relation we are after -i.e. a vertical relation that makes mental and physical causation compatible- out of a set of relations that exclude one causality or the other.

2.2. Realization

In the light of what I have just said, in order to solve the exclusion problem, what we need is a deep explicative relation of dependency that links mental and physical properties. Given that functionalism has been the dominant theory of mind for quite a long time, it is a bit odd that philosophers have not thought of the realization relation as a candidate until recently. According to functionalism, functional properties are properties individuated by their causal roles, that is, they are dispositions; however, it was

functionalism's great improvement over behaviorism to say that these causal roles have a "categorical basis", also called 'realizer'.

A functional property, furthermore, asymmetrically depends on its realizer in the sense that a particular x cannot instantiate a functional property if it does not instantiate one of its realizers, and, if any particular x instantiates a certain realizer R , then for every y that shares context and laws of nature with x , if it instantiates R , it must instantiate the functional property R realizes. That is, the realization relation, parametrized to context and laws of nature, falls under the supervenience pattern.

The interpretation of functionalism and therefore of the realization relation, however, was a matter of dispute when it was first proposed. Some (paradigmatically, David Lewis) understood it as a reductive theory, while others (as, for instance, H. Putnam) saw in it an antireductionist account. The difference lied in the account of functional predicates advanced by ones and the others: reductionists held functional predicates were nothing but descriptions of the realizer -'the property that has such and such causal powers'- whereas antireductionists claimed functional predicates referred to second order properties, being 'second order property' defined as follows: (Kim (1998); 20)

"F is a second order property over the set of base (or first order) properties B iff F is the property of having some property in B such that D(F), where D specifies a condition in the members of B". (That specification being by causal roles in the case of functional properties).

The reason behind this idea of second order properties is that these are multiply realizable, whereas first order properties aren't. As the fact that mental properties are multiply realizable was one of the strongest motivations for rejecting identity theories, the functionalist orthodoxy was easily fixed in this point: mental properties are second order functional properties related to physical properties by an anti-reductive relation of realization.

Now, can functionalism so understood be a good theory of mental causation? There are two important arguments against an affirmative answer. The first, developed specially in Block (1990) says that functional properties cannot cause the effects they are supposed to cause because they are conceptually connected to such effects: although the existence of a conceptual connection does not exclude the presence of a nomic connection, the fact is that, except in rare occasions, they do not come together. Although compelling at first glance, I would not bet on the soundness of this argument.

The second argument is both deeper and sounder (and, by the way, is also offered by Block in the same paper). It is Jaegwon Kim who has

developed it more thoroughly and made more use from it. According to this argument, realizationism is committed to what Kim has called “the principle of the causal inheritance”, namely, the fact that realized properties do not have causal powers over and above the causal powers of their realizers. That is, functionalism is committed to the view that the effects to be explained are caused only by the instantiation of the realizing properties. Realized functional properties do not make any causal difference.

Besides, if this is combined with the famous *dictum* of Alexander (to be is to have causal powers), we have that there are no such things as second order functional properties, but only second order functional concepts. This result, as Kim emphasizes, should not be that surprising, since it would be really peculiar that the simple existential quantification over properties did bring in new properties to a thus overpopulated world.

Therefore, on this account, functional properties do not exist as something distinct from their realizers; in Kim’s words, they are “nothing but” their realizers: in each instantiation, the second order property is identical to the property that realizes it. So we have the explicative relation we were after, but at the cost of a reductionism that is even worse than classical identity theories, for there the reduced properties were global whereas what we have now are *reducens* that are extremely fine-grained. In other words, multiple realization amounts to multiple reduction, and so realizationism comes to the denial of the first premise of our exclusion problem, namely, the causal efficacy of mental properties.

The lesson to be extracted, I think, is that it is not enough to say that one property depends on another, since it can be further asked: what is the relation between the causal powers of mental properties (higher level properties in general) and of physical, or lower level, properties? If the relation is of what Kim calls inheritance, like in the realization case, then mental properties are nothing but physical properties; and if there is no relation, but just difference, then we have not solved the exclusion problem at all.

3. Other candidates

The purpose of this section, then, is clear: to find a relation between mental causation and physical causation such that mental causal powers are neither distinct nor inherited from physical causal powers. That is, the causal powers mental properties bring in must be either identical or partially identical to those of physical properties, but at the same time genuine. There are two ways in which this could be possible: (a) mental properties and physical

properties are nomically related but their causal powers are independent -i.e. there is causal overdetermination, and (b) both properties and their causal powers are related. Note two things here: First, it is the contingent link between properties and powers what makes (a) possible; and second, (a) is not the usual overdetermination statement: what defenders of the overdetermination solution to the problem of exclusion normally say is that it is not necessary -not even nomically- that mental and physical *properties* coincide, though they happen to coincide when behaviors are concerned, whereas the position here is that mental and physical properties are nomically correlated; it is their powers that are independent. Now, be it as it may, I think coincidences are hard to swallow, and what Tim Crane says about property overdetermination -it commits you to the truth of statements like "I would still have gone to the cupboard for an aspirin even if I hadn't had the pain"- applies equally to our case of power overdetermination.

So let's focus on the (b) option. I want to examine two sorts of relations that may seem to fit in the picture just drawn: one is constitution and the other determination, or, more generally, genus-species.

3.1. Constitution

Constitution is a part-whole relation. It holds, typically, between a sortal and the parts that compose/instantiate the object that falls under such sortal. Thus, a dog is constituted by four legs, a tail, a head, etc. The friends of this relation distinguish two senses of identity governing it: the sortal property is, in a sense, identical to its constituent properties, since the "thick particular" that falls under the sortal and the one formed by its constituents occupy the same places at the same times. On the other hand, it is, in a stricter sense, not identical: the constituents could be absent and the sortal still present: a dog can lose his tail, or his legs be removed, and still persist as a dog.

That is the relation between sortals and constituents: what is the relation between their powers? At first glance, it seems as though there is a heavy coincidence, even partial identity: a body, for instance, seems to cause a lot of the same things the sum of its parts cause. It does not cause all the same things just because it is not strictly identical to such sum, and so the persisting powers of the body cannot be strictly identical to the powers of the sum. A body can lose its hands and still be a body; however, that loss has severely shortened its causal powers. The body, then, still exists and seems to cause things, but it has been shown that its powers are not identical to those of the sum that composed (or composes now, or will eventually compose) it. Thus,

it seems plausible to say that the persisting powers of a sortal are a subset of the powers of its constituent properties.

Now, is it really reasonable to say that a sortal has genuine persisting causal powers? It seems to me that we are in a position similar to that of realizationism: whatever causal powers -persisting or contingent- a sortal may display, they seem to come from below, so to speak. That is, I see no clear way in which property constitution, as it has been explained, can avoid Kim's inheritance principle.

3.2. *Genus-species*

The relation holding between a genus and its species, which following Aristotle and the medieval work on predicables is a relation of inclusion, looks more promising. According to this view, a species is a genus plus a difference: the property of being a man, thus, amounts to the property of being animal plus the property of being rational. This account of the genus-species relation coincides with the view on the determinable-determinate relation that has been put forward, among others, by Stephen Yablo in a couple of influential papers (1992, 1997), namely:

(#) P determines Q iff: for a thing to be P is for it to be Q, not *simpliciter*, but in a specific way.

The determination relation has some intricacies that make it unique, different from any other relation. It holds between a color and its shades, between the property *color* and the different colors, or between a certain quantitative property such as *mass* and the different values mass can take. And it seems to obey its own logic, first proposed by W.E. Johnson back in 1921. Many philosophers take this relation to be a case of the more general genus-species relation, but some others, like D. Armstrong do not. I will join those who think it is a case of the genus-species relation, and so I will go on to discuss this latter case, assuming -plausibly, I think- that the intricacies of the determination relation play no substantial role in matters of causal exclusion.

As is well known, Yablo has claimed that the determination relation thus understood can solve the mind-body causal exclusion problem. In what follows, I will examine this claim, starting by dividing it in two: (a) the determination (i.e. genus-species) relation can solve exclusion problems, and (b) mind and body are related by such relation.

The first idea to examine, then, is whether properties that stand as genera to species exclude each other in causal matters or not. As said above, this amounts to the question of whether the causal powers of a genus are (partially) identical to those of its species without having been inherited from

it. And I would say there are, at first glance, prospects for a positive answer. Let's see: Genera are postulated in order to explain similarities between species. As said at the beginning of the paper, I take it that properties have an internal structure or "nature": resemblances in that nature may well lead us to say that they resemble in virtue of having a property in common; in other words, by belonging to the same genus. So a genus becomes part of a species. Now, let's take species A and B of genus X: A and B have some causal powers in common, as do all the species of genus X: would not we say that they have such powers precisely because they have X as a part? That is, would not we explain their having causal powers in common in exactly the same way we explained their similarity? And if this is so, then, we are claiming genus X has some genuine causal powers, not inherited from below.

That's the basic reasoning, I think. First, we put genera inside species, so to speak, by means of the similarity argument. Then, we explain common causal powers by that common component of all the species. And we conclude that those causal powers belong to the genera.

It may be useful to note the contrast with realization pairs: there, realizers do not have to resemble each other but causally, that is, they may very well differ according to their internal structure. Functional groupings are made only attending to causal roles; it may very well happen that all the realizers of a given functional property resemble each other, but that is not what leads us to group them: only causal behavior does. This way, we miss step one in the argument, and we cannot independently explain common causal powers by invoking the functional property. A functional predicate and a genus predicate may be coextensional; furthermore, functional similarities may be good clues to similarities in internal nature, but that is clearly not enough to make functional properties ontologically respectable. Note finally that all this would not make sense if the view on properties I have taken is wrong and there is no such thing as an internal structure and properties are just causal powers: then, the difference I was trying to make would vanish.

Now, suppose what I say is true, and genera have causal powers of their own. Then, the following will be also true: some occurrences of, e.g. scarlet, are causally efficacious as scarlet-events, while some others are efficacious just in virtue of scarlet being a species of red. That is, we seem to confront an exclusion problem once again, this time upside down, so to speak: species, though causally efficacious at times, inherit many other times their causal powers from their genera. Thus, when the bull charges against the cape of the bullfighter its scarlet color is to be held responsible for the charge just

because it is a shade of red. It is the red-event, not the scarlet-event, what made the causal work.

It is hard to settle intuitions here. Some people have intuitions that grant privileges to all that looks very concrete, and so won't even accept my description of the problem; for them, what *really* happened in the world, the only thing that was there, was the scarlet-event, so it was that event which entered the causal relation. Some other people, on the other hand, are more moved by the counterfactual-supporting character of causal relations; from that point of view, it is the redness of the cape what caused the charge: otherwise, counterfactuals do not come out right. Yet there are those who see no exclusion problem between genera and species: exclusion problems take place between properties and powers that are different, not between properties and powers that are either identical or partially identical. My own intuitions sway between this last position and the second: According to what I have been saying, the first type of intuition can get a response, namely, that just as it is true that what was there is a scarlet-event, so is it that scarlet occurrences are red occurrences, being scarlet red plus a difference.

So, my final position is disjunctive: either there is no exclusion problem affecting two properties that are related as genus to species, given that one -the genus- is included in the other, or there is a problem, and some of the powers of the species are inherited from the genus.

Let us move then to the second part of the question, that is, whether mental properties and physical properties can be pairs of the genus-species relation. As said above, Stephen Yablo has written a couple of influential papers defending an affirmative answer to this question: mental and physical events, he says, are related as determinables to determinates. However, he does not give any argument supporting this claim: Is it so obvious that the determination relation, or something very similar links mental and physical properties? I would say it is not. If properties A and B -elephant and whale, for instance- are different, and are species of property X -mammal-, then A and B cannot be exactly similar when viewed from X. That is, they cannot be the same kind of X: otherwise, they would be the same species. I think that is obvious. However, different physical properties can be exactly the same kind of mental property, as arguments for multiple instantiability have shown. But if physical properties P and Q were decomposable into mental property M and a difference (one for each), then it could not possibly happen that P and Q were the same mental property. Of course, this is so because the way physics individuates its properties is independent from the way psychology does. (Whereas mammal identity and elephant identity are closely related matters).

A related possible attack to Yablo's claim, developed in McGrath (1998), is that the determination relation is discovered *a priori*, and there is no such *a priori* link between mental and neurological concepts/properties. I think this may be a shorter way to express what I was trying to say.

In any case, it is surprising not to find arguments in Yablo's papers. The surprise nonetheless strengthens if we understand that in his (1992) Yablo is concerned, more than anything, with narrow mental properties, whereas in his (1997) he is explicitly defending the determination relation between physical and *wide*, that is, relational, mental properties. I think it is this feeling of uneasiness what makes Kim put, in his last book, the determination attempt together with the strong and also the mereological supervenience (defended by Segal and Sober (1991)) attempts and claim that all of them lack explanatory power.

All this can be said to be unfair, for, as a matter of fact, Yablo (1992) spends a good amount of pages defending that mental *events* are determined by physical *events*. Yet, a more or less close look at what he says there should reaffirm us in our position. Mental events, he basically claims, are determinables whose determinates are physical events because the essence of mental events is included in the essence of physical events. Now, conceded that view on determination, it is legitimate to ask how come the essence of physical events contains the essence of mental events. It does seem as though physical events (specially, if we are speaking of neurological events) do not have intentional (or qualitative) properties in their essential pool, while it is essential to mental events to have intentional (or qualitative) properties.

Yablo's response to this query is quite disenchanting: he says we can be sure of mental events having their essences included in the essences of physical events because mental properties supervene on physical properties, so that if a certain property M is essential to a mental event *m*, as M depends on a certain physical property P, then, P is going to be essential to the physical event *p* that subvenes *m* and, hence, M will be also essential to *p* (given that P necessitates M). As can be imagined, I say this response is disenchanting because it makes Yablo's ingenious first claim -the determination claim- rely on the doubtful property supervenience thesis that we have discarded before.

Again, all this depends heavily on the view on properties I have taken. If causal powers were to be held (part of) the essence of properties, then what Yablo claims easily comes out true. But note also that, as said above, this second view on properties makes the determination claim no different from the realization claim.

In any case, my purpose was to see how far we could go, starting from some sensible standpoints, towards an equally sensible solution to the mind-body exclusion problem. And this is as far as we got: genus-species relations look interesting, but it is hard to see how physical properties can turn out to be species of mental properties. Once again, and as usually when the exclusion problem is concerned, a not incredibly optimistic conclusion.

REFERENCES

- Armstrong, David (1997) *A World of States of Affairs*. Cambridge: Cambridge University Press.
- Block, Ned (1990) "Can the Mind Change the World?", in Boolos, ed., *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge, Cambridge University Press.
- Johnson, W.E. (1964) *Logic: Part 1*, 3, New York: Dover Publications.
- Kim, Jaegwon (1984) "Supervenience and Supervenient Causation". *The Southern Journal of Philosophy*. Vol XXII, Supp.: 45-57.
- Kim, Jaegwon (1993) *Supervenience and Mind*. Cambridge: Cambridge University Press.
- Kim, Jaegwon (1998) *Mind in a Physical World*. Cambridge, MA.: MIT Press.
- McGrath, Matthew (1998) "Proportionality and Mental Causation: a fit?", *Philosophical Perspectives*, 12, 167-176.
- Segal, Gabriel and Sober, Elliott (1991) "The Causal Efficacy of Content". *Philosophical Studies*, 63: 1-30.
- Yablo, Stephen (1992) "Mental Causation". *The Philosophical Review*, 101: 2, 245-280.
- Yablo, Stephen (1997) "Wide Causation". *Philosophical Perspectives*, 11: 251-281.

Agustín Vicente,
 Universidad del País Vasco
 e-mail: ylbvibea@sf.ehu.es

Epistemology and Philosophy of Science

Epistemología y Filosofía de la Ciencia





FALLIBILISM AND REALISM IN LAKATOS

JESÚS ALCOLEA-BANEGAS

Universitat de València

JUAN M. LORENTE-TALLADA

Universitat de València

JOSÉ P. ÚBEDA-RIBES

Universitat de València

ABSTRACT: This contributed paper studies the Lakatos's conception of the nature and the development of mathematical knowledge and, in particular, his declared fallibilism. The aim is to expose some realistic elements of his work, analyse their compatibility with his fallibilism, and characterise the resulting philosophy as a fallibilist critic realism.

KEYWORDS: Lakatos, mathematics, quasi-empiricism, fallibilism, realism, criticism.

The thought of I. Lakatos on the nature of mathematics and mathematical knowledge was developed in two stages. In the first one, represented in particular by *Proofs and refutations*,¹ Lakatos defended a Popperian conception of the nature and the growth of mathematical knowledge, in accordance with which mathematical theories are fallible and quasi-empirical, and mathematics progresses formulating bold conjectures, which remain subjected to criticism and refutation. The second stage is characterised by what Lakatos called "the methodology of scientific research programmes", in accordance with which, scientists work in a comprehensive frame of accepted hypothesis and methodological suppositions called "scientific research programmes". Programmes are rarely refuted and are superseded by competitive programmes of greatest "success". In the frame of this presentation of Lakatos's thought, our aim is to expose some realistic

¹ With the abbreviations *PR*, *MSE*, and *MSRP*, we will refer to the works by Lakatos, *Proofs and refutations* (1976), *Mathematics, science and epistemology* (1978), and *The methodology of scientific research programmes* (1978), respectively, published in Cambridge by Cambridge University Press.

elements of his work to the view,² analyse their compatibility with his declared fallibilism, and characterise the resulting philosophy.

As it is known, the main reasons that could be offered in favour of the mathematical realism are: (1) The success of the mathematical theories, that could not have been produced by chance. (2) The persistence of the mathematical theories along the time. (3) The own statements by some mathematicians in the sense that they are dealing with something real. However, against the realistic position, and from a Lakatosian perspective, it has been put forward: (1) The permanent refutation that mathematical theories, conjectures and proofs have to undergo in order to be developed; i.e, the provisional, corrigible, and fallible character of mathematical knowledge. (2) The mathematical propositions (proofs) that once were thought to be true (correct) and that later turned out to be false (incorrect). In our view, Lakatos's methodology and epistemology are supplemented with a realistic ontological position. But, maybe under the influence of the ideas of Pólya and Popper, it does not mean that in this respect his theses are affirmed with the same clarity and radicality that his epistemological theses. However, the realistic ideas of Popper and, mainly, his theory of the third world could light up the way to follow with Lakatos. We seek then to detect those elements that allow us to characterise Lakatos's conception, elaborating it and defending it like a certain type of realism, next to the Popperian version of the third world, but with an especially conjectural element.

Lakatos becomes interested for the search of the rationality in the history of science and of mathematics. In this way, he is centred in a rational reconstruction of the life of the mathematical or scientific thought. It could be said that a theory or a conjecture is replaced by another logically better. The rationality that guides the science is the logical method. Nevertheless, it could not be proved that the logical method takes us unavoidably to the truth. In fact, it could be supposed that progress in mathematics means that the concepts used by the mathematicians approach in an increasingly narrow way to the objective structure of mathematical reality. So, Lakatos says: "*Naive conjectures and naive concepts are superseded by improved conjectures (theorems) and concepts (proof-generated or theoretical concepts) growing out of the method of proofs and refutations.*" (PR 91) So that "*I find it uninteresting to quarrel not only about the exactness of naive concepts but also about the truth or falsehood of naive conjectures*" (PR 90), because

² The topic of realism in Lakatos has been approached by other authors. Only as a reference, and without us being able to enter to discuss their positions, we will mention: M. Steiner: "The philosophy of mathematics of Imre Lakatos", *Journal of philosophy*, 80 (1983), pp. 502-521; J. R. Brown: "Proof and truth in Lakatos's masterpiece", *International studies in the philosophy of science*, 4 (1990), pp. 117-130; and D. Corfield: "Assaying Lakatos's philosophy of mathematics", *Studies in history and philosophy of science*, 28 (1997), pp. 99-121.

knowledge has to be developed, and with this development the language changes. However, “in the end we shall arrive from naive, accidental, merely nominal classification to the final true, real, classification, to perfect language!” (PR 92) That is, with the method of proofs and refutations the theory is developed and “the balance changes in favour of the realist.” (PR 92, note 1) The history begins with a problem for which there is an answer in a primitive conjecture, that remains subjected to criticism, through which the solution improves and eventually the final result arises. Along with the process it is emphasised the indispensable character of the language to argue and discuss critically. Criticism carries out the role of forcing a displacement toward a new conceptual frame in search of deeper theorems and proofs, but it could not be proved that the method takes us unavoidable to the truth. In fact, Lakatos does not hesitate to prove even a false conjecture through its decomposition in subconjectures. The truth and the falsehood are in the same level with the method of proofs and refutations. The difference between them is that the increase of knowledge implies a reduction of the falsehood by means of a counterexample and an increase of the truth by means of a proof. So, although his fallibilism follows from his insistence in that the refutations always imply new conceptual development, the logic of mathematical discovery takes us more and more close to the truth. Now then, upon talking about the mathematical truth, we are talking about the “correspondence with the mathematical facts” and something more, something that does not reside in us, since the relation of correspondence could be the case without our knowing it. These considerations, based on the Tarski’s theory of truth, took Popper to a theory of objective knowledge, of the knowledge that exists outside of us and, in part, independently of us. This theory takes the form of a pluralistic metaphysics of the three worlds, that Lakatos seems to subscribe.³ As he says, “The *first* world is the material world, the *second* is the world of consciousness, the *third* is the world of propositions, truth, standards: the world of objective knowledge”. (MSRP 92, note 2; cf. MSE 108) Lakatos acknowledges the importance of the third world in order to understand in what sense the science is objective: “*The – rationally reconstructed – growth of science takes place essentially in the world of ideas, in Plato’s and Popper’s ‘third world’, in the world of articulated knowledge which is independent of*

³ In speaking about debts or influences between Lakatos and Popper, it would be a good thing if we bore in mind the following statements by the last one: “The main point here I owe to Lakatos’s philosophy of mathematics. It is that mathematics (and not only the natural numbers) grows through the criticism of guesses, and bold informal proofs. This presupposes the linguistic formulation of these guesses and proofs, and thus their status in the third world. Language, at first merely a means of communicating descriptions of prelinguistic objects, becomes thereby an essential part of the scientific enterprise, even in mathematics, which in its turn becomes part of the third world.” (*Objective knowledge. An evolutionary approach*. Oxford: Clarendon Press, 1972, p. 136.)

knowing subjects. (...) But the mirror-image of the third world in the mind of the individual – even in the mind of the ‘normal’ – scientists is usually a caricature of the original; and to describe this caricature without relating it to the third-world original might well result in a caricature of a caricature. One cannot understand the history of science without taking into account the interaction of the three worlds”, (*MSRP* 92-93) which enjoy autonomy. The necessity of the third world is given expression in the necessity of separating the very argument of the beliefs and desires of the participants in the dialogue, because it is necessary to show the rationality of the process and it could only be done identifying after the successive improvement the underlying argument, that lives and is developed in the third world together with propositions, theories, problems, problemshifts, research programmes, etc. Lakatos’s sympathy for the thesis of the third world is shown in his criticism of elitism and pragmatism (*MCE* 112ff). The first one evaluates the objects of the second world, instead of the objects of the third. The second one does not distinguish between rational belief (third world) and practical belief (second world) and it abandons the conception of the objective truth, that is a matter of the third world.

Although Lakatos thinks that “the bulk of logic and mathematics is God’s doing”, he declares himself as “a fallibilist not only in science, but in mathematics and logic as well” (*MSE* 127), because there is not definitive proof that we have been advancing toward the truth and “we can only (non-rationally) believe, or, rather hope, that we have been.” (*MSE* 223) We can understand then that mathematics is part of the third world, which we discover it, but that we only know things of that world in a fallible way. Lakatos separates the life and the social context of the mathematicians from the logical interaction of their abstract objective ideas. That realm of ideas is the consequence of the logic of mathematical discovery, of the human activity that produces it, and from which mathematics is alienated, acquiring its autonomy (Cf. *PR* 146). It contains the germ of an idea relative to the nature of mathematical objects, that we are able to conjecture as if they were created, and acquiring certain autonomy, so that they could be discovered. *PR* underlines that in mathematics, and concretely in the mathematical activity, we attended a constant expansion of the domain of the facts, of the objects and of the mathematical methods. So, although for the realist the aim lies in articulating true propositions, because the core of science and mathematics is constituted of true propositions, for Lakatos, the mathematical activity is previous to the emission of conjectures and, therefore, previous to the problem of the truth of mathematical propositions. From the point of view of the mathematical activity, avoiding the errors and criticising the theories and the false conjectures are as important as looking for the truth. However, the essential point is not the truth, but the innovative and transforming capacity of the activity. This capacity has to do with the improvement of the concepts, the

conjectures, the theories and the methods, but also with the discovery of new mathematical facts, what constitutes a good argument in favour of realism. The idea of the discovery of the mathematical objects could be explained through the Popperian conception: There are certain things about those objects that we have not carried out, but that we could discover. An important aspect of that conception, and that it marks the difference between the objects of the third world, in general, and the mathematical objects, in particular, is that we could recreate these objects.⁴ This recreation could be understood in the light of the Popperian perspective because the mathematical objects depend on the second world (of our mind), and in the light of the Lakatosian perspective through the dialectics of the proofs and refutations, of the improvement of our conjectures toward a better outline of what we know about those objects and in this sense we recreate them after their having been created or discovered. For Popper, we create the natural numbers; the rest is things that we discovered of them, their properties, etc. Quoted by Lakatos, the discoverer of the proof-generated concept of uniform convergence and of the method of proofs and refutations, the mathematician P. L. Seidel (*PR* 136), says that we discover that “the theorem cannot be generally valid.” (*MSE* 47, note 1) That is to say, we discover things of the numbers, and we express them by means of propositions, that sometimes are proved. Nevertheless, we also discover that our theorems are not valid or that our proofs need improvement after the criticism. The image that provides Lakatos with the heuristics allows us to create and recreate the objects. But, at the same time, we recreate our knowledge and we improve it thanks to criticism. Conjectures arise starting from problems, like rational answers to them, while problems arise with critical discussion. Explaining the emergence of a particular conjecture is equal to relate it with a specific problem, and with the state of the discussion, that gives rise to that problem. But this third-world relation between problems, conjectures and states of the discussion has no similarity with the causal relations as we found them in the first world. The third-world relations are rational relations, such like the presupposition, the implication and the appropriate character or relevance of an answer to a particular question in a particular way. Then the history and the philosophy of mathematics have to do with rational beings, that have thoughts, that pose and solve problems, formulate conjectures, draw consequences, etc. These thoughts, problems, conjectures, etc., are members of the third world, and the historian and the philosopher are interested in establishing their third-world relations, by means of rational reconstruction. So, the diversity of conjectures, tests and strategies of inquiry is a consequence of the fallibilism and has a critical function.

⁴ K. R. Popper and J. C. Eccles: *The self and its brain*. Berlin: Springer, 1977, p. 44. See also the work by Popper quoted in the previous note.

Lakatos is really interested in what he calls “the full empirical or quasi-empirical stuff” that it is the object of the informal theory. In this sense, he states precisely that the pre-formal and post-formal proofs “prove something about that sometimes clear and empirical, sometimes vague and ‘quasi-empirical’ stuff, which is the real though rather evasive subject of mathematics. This sort of proof is always liable to some uncertainty on account of hitherto unthought-of possibilities. [The formal] proof is absolutely reliable; it is a pity that it is not quite certain – although it is approximately certain – what it is reliable about.” (MSE 69) We can guess that that “stuff” is that world of mathematical objects, formed by their interactions and relations, which we could only know in an approximate way. In being abstract objects we could only know them through their properties and relations, and it is exactly this that we expressed by means of our conjectures. We play nominally with them and their properties in a critical process of refutation and improvement, but the approximation leads us to know them better, although that knowledge is never definitive. In the theories that we formed on that world, Lakatos shows his preference for the informal ones, for their wealth, and without disregarding the formal ones. In fact, he clears up (MCE 60) the transitory character *also* of these theories, whose enormous heuristic power we can suppose, but he acknowledges the rigour of the formal proof in those theories. Nevertheless, nothing is definitive and theories are always in growth and there is no form of predicting how they could change their relative force. The caution of Lakatos at the moment of dealing with the formalized axiomatic systems is emphasised in the question that he asks (MCE 66) on the impossibility of obtaining a counterexample in a formalized system, supposedly consistent, in which the theorem of Euler on regular polyhedra is proven. The problem is that there is no guarantee that the system contains the stuff with which we are dealing in the informal theory, and there is no formal criterion with respect to the correction of the formalization. It could be that the formal theory were infallible, but in it “imagination is tied down” (MSE 68), while in an informal theory, because it is open, we could always introduce new terms, axioms, rules, etc., whether it is or not in a fallible way. Lakatos gives examples even of axiomatic theories, like the theory of probability, that they have needed certain modifications.

Although some mathematicians admit their metaphysical leanings and the importance of these for the growth of their mathematics, Lakatos points out that there is a tendency “to demarcate in mathematical theories a *hard formal core* which is uncontested and indubitable, eternal; and a ‘*metaphysical*’ interpretation of the formalism which is controversial, ‘soft’, changing.” (MSE 58) However, “one cannot separate in informal [and in formal] mathematics ‘metaphysics’ from ‘technicality’.” (MSE 59) And this happens in the growth of the informal mathematics taking its characteristic pattern into consideration that, like method of mathematics, it systematises the

method of learning from our errors and that it is summarised in these four points: (1) We formulate a (primitive) conjecture, as a result of the study of a *problem*, supported by an experimental investigation. (2) We try to prove the conjecture making use of a mental experiment or approximate argument. (3) Through the critical discussion of the proof and of the conjecture, the knowledge is developed by means of the elimination of some errors, so that it is learned how to understand (the problem and) the original conjecture, as well as the necessity of a new proof. (4) The critical discussion of our better conjectures always brings to light new proofs, improved versions of the original conjecture and new problems. This characteristic pattern “shows that the standard procedure of lemma incorporation carries the theoretical concepts of the proof into the theorem: improvement by lemma-incorporation means theory-impregnation.” (*MSE* 59) Those aspects of mathematics with which the reader of a work is supposed to be familiarised in the starting point (lemmas, true or false, that are considered so trivial as not to be worthwhile to mention them, for instance), are decreasing “as criticism turns background knowledge into knowledge,” because “knowledge has no foundations”. (*PR* 45-46, note 3) In this line, there are conscious suppositions that have to be made explicit in the language in order to be put to criticism. Because through the analysis of proof, criticism allows us to discover the error, the truth, the hidden lemmas and the hidden conditions to the proof and to the theorem (*PR* 46). “It is these hidden lemmas that contain sophistication and fallibility and ultimately destroy the myth of infallible deduction.” (*PR* 81)

The analysis of the process of critical discussion of conjectures makes clear that the discussion never considers the question if a conjecture is “justified” in the sense that *we* are justified to accept it as *true*. At the very best, the critical discussion justifies the statement that the conjecture in question is the best that it is provided, or, in other words, that it is the one that comes nearer to the truth. The process of critical discussion is a process of rational discussion. The participants are willing to discuss critically their own beliefs and to correct them in the light of the critical discussions with other people. It is a principle of minimum rationality the one that encourages our conjectures and our proofs and although we know that our conjecture is not true, we have some reason in order to consider it as a good approximation to the truth.⁵ So, the state of the mathematics is unavoidably hypothetical, and

⁵ If the progress is conceived as a gradual approximation to the truth, we arrived to the concept of verisimilitude. We know about the problems that Popper had with this concept, mainly at the moment of characterising the verisimilitude of two theories. However, we believe that in Lakatos’s case –critic of Popper in this point– it is possible to avoid the problem because one refers only to the verisimilitude of mathematical conjectures, and in this case the verisimilitude depends on the number of truths and of falsehoods that this conjecture implies, since the conjectures, the counterexamples and the anomalies are evaluated critically by their consequences.

the mathematics of the moment describes the mathematical reality in an approximate way. There is no reason to suppose that that description is trustworthy. The thesis that the mathematics describes the real mathematical world should be considered more as a question of intention than as an accomplished fact. Sure that we are not in conditions of affirming that mathematics, as it is, reaches a characterisation of the mathematical reality. In spite of everything, the mathematicians continue having realistic leanings. So, Y. Moschovakis talks about the strong realist feelings of the mathematicians and of how “they spend their working hours behaving as if they were completely unabashed realists”, and J. Dieudonné points out that the mathematicians also have a feeling that they are working with something real.⁶ To this respect, the Gödel’s position is very important. He claims *clarity* to the general mathematical concepts “for us to be able to recognise their soundness and the truth of the axioms concerning them”, which sometimes “force themselves upon us as being true”. This is a sign of the action of the third world on the second world. Now, how do we obtain that clarity and correction, if it is not through the method of proofs and refutations, method that deprives us of the error and does it help us to understand? The consequence would be that “the set-theoretical concepts and theorems describe some well-determined reality”, although the system would always be incomplete and would need of the addition of new axioms.⁷ The *aim* of mathematics is, without a doubt, to answer to our questions about the mathematical world *correctly* and to describe that world “as it is in reality”. For Lakatos, as for Gödel, the orientation of the mathematics is objective, but also factual, because it takes care of establishing *true* facts on the real mathematical world. In Lakatos case in a fallible way.

The language of mathematics supposes a descriptive commitment. In the semantic level, mathematics makes statements about how the things of the mathematical world are, so that the realism is connatural. This status factually descriptive is founded as a last resort in the continuity of their problems with those of the daily life (first world). We begin with problems, and we put an end to problems, as the Teacher says citing Popper in *PR*. Mathematical tools are supposed to allow us to solve problems on the first world. The third world acts on the first world through us (second world). That’s why, mathematics does not refer to a realm totally independent. Although we consider that mathematics describes *really* a mathematical reality, there is no way of knowing *really* that the mathematical theory provides descriptively correct

⁶ Y. Moschovakis: *Descriptive set theory*. Amsterdam: North-Holland, 1980, pp. 605-606; and J. Dieudonné: “The work of Nicholas Bourbaki”, *The American mathematical monthly*, 77 (1970), p. 145.

⁷ K. Gödel: “What is Cantor’s continuum problem?”. In *Collected works*. Vol. II. Edited by S. Feferman *et alii*. New York: Oxford University Press, 1990, pp. 258, 268, and 260, respectively.

information about the mathematical world, and we should conform ourselves with understanding it as the representation of our best estimation of how is the mathematical reality, since it will always be incomplete. We are not, then, in the face of a platonistic realism of Gödelian style. Lakatos rejects the essentialism inherent in this type of realism, that demands that proofs, explanations or putative definitions are final and certain. So, the objective criticism by means of counterexamples (PR 16) requires of a previous agreement on the meaning of the terms and for them some definitions are proposed. But there is nothing like a definition that it is *the* right definition, because the definition has also a conjectural character and it is not something that captures “the essence of the concept involved”. (PR 123) In fact, the counterexample leads to revise the definition in order to free it from counterexample. Therefore, we are not justified in order to claim the definitive mathematical truth, neither the upholding or final correction of the mathematical statements, because the engine of criticism does not stop and, as it changes, it leads to the truth and “nothing is settled” (PR 105 and note 1). With his statements the mathematician assumes a descriptive commitment with the real mathematical world, but the provisional character with which those statements are formulated in order to subject them to criticism and the improvement introduces a conjectural and fallibilist element that allows us to present Lakatos’s position as a kind of fallibilist realism. We recognise the reality (not totally independent of the mind) of a real mathematical world, that we could only know in an imperfect and approximate way, and we recognise that, independently of what we are able to extend the border of mathematics, it always remains more to be done, to be guessed. The mathematical reality surpasses our cognitive grasp. There is more in the mathematical world of what there is to our grasp. The consequence is that, for Lakatos, (informal) mathematical theories are empirical or quasi-empirical theories and, so, although they could be well corroborated, they will never be true and they will always be conjectural. Lakatos adheres to Pólya’s slogan, “You have to guess a mathematical theorem, before you prove it,” and to Popper’s, “We never know, we only guess”. In fact, what matters is not how it is known but how our conjectures are corrected and improved through critical discussion. Thus, if the theory is not shown to be false, it “will remain conjectural for ever”. (MCE 19) For Lakatos, Gödel’s first incompleteness theorem is a principle of conservation of the fallibility. “But this need not make us give in to mathematical scepticism: we only have to admit the fallibility of daring speculation. Gentzen’s consistency proof – and Tarski’s semantic results – are real [victories] (...) even if one admits (...) the definitely conjectural character of the new methods,” in such a way that one can “honestly admit mathematical fallibility, and try to defend the dignity of *fallible* knowledge...” (MSE 22-23) In fact, the mathematicians do not worry about if their arguments could be expressed in any formal system, but of if they are valid

ways for reaching the truth. Therefore, Gödel's theorem shows that mathematical truth is beyond all formalism. Many mathematicians claim for intuitively convincing methods, because, as Lakatos says (*MSE* 23, note 1), "if one gives up intuition, despairs of certainty, and equates knowledge with certainty, then one may turn one's back on truth and play around with formal systems..." That is, we could not give up to the intuition, but its fruits have to be subjected to the tribunal of criticism, since all intuition is fallible.

For what we have said, it seems plausible to uphold that the realism is perfectly adapted to what is and to what should be the mathematical activity from Lakatos's perspective. The problem would reside in what variety of realism would be possible to choose as adequate interpretation of that activity. We opt for a *fallibilist critic realism*, because it is centred in a critical and active position of the individual, that involves the revision of the knowledge and the improvement of the conjectures and, therefore, of the theories, and in an ontology where the objects, fruit partly of the human activity, enjoy independence from the human mind and permit an objective and gradual knowledge. This realism admits the difficulty of recognising the truth, so that the mathematical progress always lies in coming near the truth and in valuing rationally the progress of knowledge toward that truth. We would like to synthesise this realism in the following theses:

(1) Mathematics is a fruit of a human activity that has the truth as an essential aim without there being security of the success in its achievement. At any rate, the best explanation for its practical success is to assume that the mathematical theories are approximately true because they are made up of conjectures filtered by the critical method of proofs and refutations, a method supposedly rational and progressive. The authoritarianism in mathematics is bound to the idea of proving the conjectures in a definitive way. The critical approach is bound to the idea of trying to refute them.

(2) The historical factors intervene in the dynamic character of the mathematical activity, but they are not incompatible with the existence of the mathematical objectivity and the approximation to the truth, because the critical attitude allows us to detect the error and the anomalies, and to explain why a conjecture or a proof has been replaced or improved by another.

(3) The mathematical objectivity is not another thing that the critical approach. That is, the mathematician should submit the conjecture to his own and other people's criticism with the purpose of finding a possible refutation. Without criticism could be lost the autonomy of mathematics and its self-corrective character. That's why, the social or public character of mathematics is the one that safeguards the objectivity of mathematics and its tradition of discussing critically the new ideas.

(4) The truth is a regulative idea (Popper) that guides the mathematical research. It is a relation between language and reality whose meaning remains explained by the conception of truth as correspondence and it is captured in

the systematic research that uses the method of proofs and refutations. Although the conjectures do not allow us to accede easily to the truth, that method does allow us to come near the truth and to value rationally the progress in such a sense.

(5) Our knowledge is conjectural and, therefore, fallible and corrigible, and it is of a third world of objects that we go knowing from a diffuse form, the bold conjecture, the introduction of naive concepts, through proofs and refutations toward a clearer position. (The origin could be guessed to be the change or exchange, the computation, as Popper supposed.)

(6) Mathematical theories are conjectural. We could refine them with the axiomatization, but this does not remain closed because it is fed by empirical stuff and it possesses a relative reliability.

(7) The image of the progress as an accumulation of truths is rejected. Progress is considered as an approximation to the truth, without it there is a definitive mechanism that leads necessarily to a position nearer to the truth.⁸

Jesús Alcolea-Banegas
Universitat de València
e-mail: Jesus.Alcolea@uv.es

Juan M. Lorente Tallada
Universitat de València
e-mail: Juan.Lorente@uv.es

José P. Úbeda Ribes
Universitat de València
e-mail: Jose.P.Ubeda@uv.es

⁸ Our thanks go to Alex Lindsay for cheking and improving our English.



KNOWLEDGE AND RELATIVITY

PETER BAUMANN
Göttingen University

ABSTRACT: When we say that some person knows something, we usually think we report a fact which holds no matter how we look at and describe the relevant facts. I argue here that, at least in many cases, it rather depends on our description of the situation whether S does or does not know that p. In many cases we can see the supposed knower as either just lucky and not knowing that p or as not just luckily believing that p and hence as knowing that p. I end by arguing that all this does not lead to relativism.

KEYWORDS: Knowledge, reliabilism, contextualism, relativity to description, relativism.

When we say about another person that she knows that something is the case, we usually mean to report a fact concerning the epistemic state of this person. We might even say it is an "objective fact" that exists "independently from us". This talk about "objectivity" and "independent existence", however, does not help very much. What we mean is that the epistemic state of affairs we report holds no matter what we think, want or do. The truth about the epistemic state of a person P – about her knowledge -- does not depend on what the person A who attributes knowledge to P thinks. In this respect, the fact that Henry knows there is milk in the fridge is in the same epistemic boat as the fact that the earth has only one moon.

All this might appear completely trivial. However, I am going to argue that it is not trivial but false. Whether a person P knows that this or that is the case often depends on the context. In particular, it depends on how we, the attributors of knowledge, describe the situation P finds herself in. Knowledge is relative to description. This thesis obviously has far-reaching implications for our view of knowledge. So, what exactly does it mean? And what is the argument for it?

I. Knowledge and Reliability

Let me start with some remarks about the concept of knowledge. What is knowledge, anyway? Well, as we all know, knowledge that p requires a true

belief that *p*. We also know that this is not enough. We need to exclude the possibility that the person acquires a true belief by sheer luck (for instance, by first guessing the truth and then believing what she guessed). Accidentally true belief is not knowledge.¹ So, what makes true belief knowledge? I take it that reliabilism gives the right answer: Knowledge that *p* requires that the true belief that *p* has been acquired in a reliable way.² Since I do have my doubts about the prospects of reductively defining the concept of knowledge (just look at the debate after Gettier!), I do not say that knowledge *is* reliable true belief but rather that it *requires* reliable true belief. And this is enough for my purposes here.

Since the argument for the context-dependence of knowledge very much depends on the concept of reliability, I will say some words about it. The conception of reliability I use here is general enough to cover (almost) all different views of knowledge, internalist as well as externalist ones. All of them are, in a basic way, reliabilist conceptions. So, what I am going to say will not only concern those views of knowledge that are usually called "reliabilist" but any view of knowledge. Now, what is reliability?

Let us call the way a person acquires a belief her "method" of belief-acquisition (in short: "M"). It is not necessary that the person uses the method intentionally or consciously (cf. the case of perception). What is necessary is, of course, that the person uses the method correctly (and I will only talk about correct uses here). Knowledge requires that the method is reliable. This means that the use of the method probably leads to a true belief. More precisely, the conditional probability that the person acquires a true belief, given that she uses method *M*, is high, that is, above a certain value *a*. Let "*m*" stand for "The person correctly uses *M*" and "*t*" for "The Person acquires a true belief". We can then put it this way:

$$(1) \quad P(t/m) > a.^3$$

The higher *P* is, the higher the reliability of *M*. It is unrealistic to expect that *P* could ever be 1; no method is infallible.⁴ Reliability allows for error but requires that $P(t/m) > a$.

(1) does not yet capture the whole idea of reliability. It is also necessary that without the use of *M* the chances of attaining a true belief would – *ceteris paribus* – be much lower, below *a*. The use of the method is supposed to make a difference:

$$(2) \quad P(t/m) > a > P(t).^5$$

¹ cf. Unger 1970, 114f.; Williams 1978, 44f.

² cf. Goldman 1986, 103 and, as an origin of reliabilism, Ramsey 1990, 110.

³ Let "*Bp*" stand for "The person believes that *p*". We can then, alternatively, say: $p \square [P(Bp/m) > a]$.

⁴ cf. Williams 1978, 33ff.

⁵ Alternatively: $p \square [P(Bp/m) > a > P(Bp)]$.

(2) captures the basic idea about reliability. Given these explanations I can now proceed to the argument for my thesis. It has to do with epistemic luck.

II. Relativity: Epistemic Luck, Good and Bad

Let us start with a simple example. Bert wants to know how late it is. He looks at a public clock which indicates 12 o' clock. Bert, however, does not know that the clock does not work and indicates the wrong time. He thus acquires a false belief. Since looking at a public clock is a reliable way to find out the time, he acquires his belief in a reliable way (as I said: reliability allows for error). We can say that there is a "defeating" circumstance C1 here (the clock's not working) which leads to a false belief even though the method is reliable. Let "c1" stand for "C1 holds". Then we can characterize such cases by adding (3) to (2) and saying that there is a circumstance C1 such that

$$(3) \quad P(t/[m \ \& \ c1]) < a.$$

Let us go back to our simple example and modify it. What if the clock accidentally indicates the right time at the moment when Bert looks at it? What if Bert at noon looks at a non-working clock that happens to indicate 12 o' clock?⁶ Bert then acquires a true belief by using a reliable method. One might say that this should be enough for knowledge. But it seems clear that he still does not know that it is 12 o' clock. Why not?

Let us look at Bert's situation. I have already said that there is circumstance C1 (the clock's not working) which lowers the chances of acquiring a true belief (about the time) by using a reliable method (looking at a public clock) below a . Now, our modification brings in another circumstance C2 (the person's looking at a non-working clock at the time the clock indicates) which raises the chances of attaining a true belief (about the time) above a again. There is a circumstance C2 such that in addition to the conditions above the following is true (letting "c2" stand for "C2 holds"):

$$(4) \quad P(t/[m \ \& \ c1 \ \& \ c2]) > a.$$

In other words, we first had an unlucky circumstance which usually supports false beliefs: The clock does not work. But now we also have a second, lucky circumstance which overrides the unlucky first circumstance: The clock accidentally indicates the right time. Such circumstances usually support true beliefs. A defeater (C1) has been defeated (by C2).⁷ There is a double coincidence here: Usually public clocks work but if they do not work they almost always indicate the wrong time:

$$(5) \quad P(c1) \text{ is low,}$$

$$(6) \quad P(c2) \text{ is low.}$$

⁶ The example is taken from Russell 1948, 98.

⁷ cf. Harman 1968, 172.

This double coincidence is lucky because (4) holds: Despite all odds, the person get at a true belief (without even noticing how lucky she is). The person has got epistemic luck, first bad luck, then good luck in the end.⁸ We now have an answer to the question why Bert, in the modified example, does not know the time even though he has a reliable true belief: There is this lucky double coincidence.⁹ In other words, there are circumstances (C1, C2) such that conditions (3)-(6) hold.

This leads me to the crucial step of the argument. One might be tempted to add here that knowledge requires reliable true belief *plus* the absence of a lucky double coincidence. But this "normality" condition won't help. Why not? The reason is that in many situations there are accidental circumstances which render the acquisition of a true belief improbable as well as accidental circumstances which render the acquisition of a true belief more probable in the end (cf. the example above). The crucial point is that it depends on our description of the situation which circumstances we focus upon as relevant and which circumstances we ignore as irrelevant. There is no fact of the matter as to which circumstances have to be considered to be the relevant ones and which the irrelevant ones. All this implies that we can describe the same situation in different ways: as containing a lucky double coincidence or as not containing one. And there is no way to tell which description is the true and adequate one. But if that is true, then it depends on our description of the facts and not on the facts alone whether a reliable true belief contains a lucky double coincidence or not. Thus, it depends on our description and not on the facts whether S knows that p or whether S does not know that p (because of epistemic luck). The same person may be said to know that p, given one description, and not said to know that p, given another description. This concludes the argument for my thesis. I do not want to claim that all knowledge bears this relativity to description; it suffices to say that at least many cases of knowledge are of this type.¹⁰

This is a very short version of the argument. So, let me try to make the argument more clear and plausible by looking at another clock-example. Suppose the clock is not working and indicates 2 minutes after noon. At this time, unsuspecting Ann looks at the clock and acquires the true belief that it is 2 minutes after noon. Under these circumstances we would tend to say that Ann does not know that it is 2 minutes after noon. One minute later, at 3 minutes after noon, the repair service arrives and makes the not yet working clock indicate the then correct time of 3 minutes after noon. At the same time,

⁸ On epistemic luck cf. Foley 1987, 174ff.

⁹ cf. for a similar argument: Zagzebski 1994, 65ff.

¹⁰ It is, of course, presupposed that we are excluding false descriptions here from the start. The thesis, then, is that even then there is more than one description that can be accepted as adequate.

unsuspecting Bert arrives on the scene and acquires the true belief that it is just 3 minutes after noon. Does he know that? Still one minute later, at 4 minutes after noon, the repair service makes the clock indicate 4 minutes after noon and – by removing a grain of dust – makes the clock work again. At the same time, unsuspecting Carol looks at the clock and acquires the true belief that it is 4 minutes after noon. Does she know that?

Consider Carol's case. We can say that she profited from a lucky double coincidence: Usually, it takes much more time for the repair service to do its work but today they have luckily found the responsible grain of dust at the very beginning. If we focus upon this circumstance – the pre-history of Carol's situation – in our description, then we would tend to rather say that Carol does not know that it is 4 minutes after noon. But we need not focus upon this lucky double coincidence; we could even ignore it. Instead, we could rather focus exclusively upon the time after 4 minutes after noon. In this case we would rather tend to say that Carol knows that it is 4 minutes after noon. Isn't the clock working properly? And isn't she using one of the best methods for finding out the time? We can say similar things about Bert and Ann (those who looked at the clock a bit earlier). If this is true, then it depends on the way we describe the facts and not on the facts alone whether there is a lucky double coincidence and no knowledge *or* knowledge and no lucky double coincidence. At least very often, we are free to choose among several descriptions and there is not the one and only adequate description. Knowledge is relative to description.

One might object here that one should choose the *complete* description of the situation. But what is this supposed to be? What belongs into a complete description of the situation – everything? Descriptions of situations seem open-ended. This is not just because human beings have finite intellectual powers. There is also a principled reason: For every description there is a more detailed one. The idea of completeness does not even make sense here. Asking for a complete description of a situation would be like expecting the class of odd numbers to be finite and asking what the largest odd number is.

It also does not help to say that we should choose the most complete description possible under the circumstances. We usually cannot compare the degree of completeness of two different descriptions without using standards of relevance: one relevant information counts more than many irrelevant ones. Standards of relevance, however, usually vary with descriptions. Hence, usually there is no way to compare descriptions as to their relative degree of

completeness. What counts as relevant information depends itself on the description chosen.¹¹

Finally, there is one more problem for the idea of a complete description of the situation. Since we – given our finite intellectual capacities – could never have a complete description, we also could not to justify our belief that our description is "on the right tracks", that is, leads to the same conclusions as the complete one. We could never be justified in assuming that there isn't an undetected lucky double coincidence lurking in the background. Thus, we could also never have enough reason to believe that another person "really" knows this or that. But the facts are different. At least sometimes, we are able to justify our belief that somebody knows something (given a certain description of the situation). Hence, *our* concept of knowledge is not "built" in such a way that the possibility of a complete description is presupposed.

III. Implications: Anything Goes?

The examples above focus on the temporal aspect of the situation. There are analogous considerations for the spatial aspects which I do not need to go into here. Let me rather resume the general upshot of my argument. At least very often, it depends on the way we describe the situation and not on the situation itself – which allows for different descriptions -- whether we should say that there is a lucky double coincidence in the situation or not. Thus, it also depends on the way we happen to describe the situation whether a given reliable true belief constitutes knowledge or not. In many cases the correct answer to the question "Does S know that p?" is not fully determined by "objective" facts (which are independent of description) There is a fundamental underdetermination here: Knowing that another person S knows that p is underdetermined by the epistemic facts about S. Only if we have a description of the situation can we even pose or answer the question whether S knows that p. Knowledge is relative to the context of the attributor's description. Given this description, S might know that p, whereas given that description, S might not know that p. This is a remarkable aspect of knowledge.

But how should we interpret this conclusion? Doesn't all this tell us that "anything goes"? Doesn't it imply that the answer to the question whether, e.g., Ernie knows there are still cookies in the box is subjective and, in this sense, completely arbitrary? Doesn't all this lead to a relativism according to which a sentence like

Ernie knows that there are cookies in the box

¹¹ Only in special cases – e.g., when one description contains one sentence more than the other description but apart from that is identical to the other one – might a comparison be possible. It is enough here to see that most cases are not like that.

might be true for me (given my description) and false for you (given your description)? If this turns out to be the case, then there is no reason left to say that a sentence like

Ernie knows that there are cookies in the box
has a truth value at all. "True for me" and "true for you" is something very different from "true". Thus, we would end up in complete scepticism with regard to our talk about other persons' knowledge. One might even suspect that this scepticism can be extended to other types of states of a person besides knowledge.

I take it that this conclusion would rather be a *reductio ad absurdum* of my argument. We are justified in taking sentences about other peoples' knowledge to be true or false. We mean to say something about the world – and not about us -- when we say that another person knows that p. If I cannot explain all this, then all the worse for my argument.

Fortunately, it does not lead to the relativist's or scepticist's hell. Given what I have said here, it is not the case that "anything goes". On the contrary: Sentences like

Ernie knows that there are cookies in the box
remain true or false. One only needs to add that there is a choice of a description involved here (usually implicitly). Given a certain description, it depends only on the facts whether Ernie knows that there are cookies in the box. Sentences like

Ernie knows that there are cookies in the box
remain true or false. The contextualism I propose here has no relativist oder scepticist implications at all. There is no danger of losing objectivity, truth or even the world.¹² The account of knowledge I argue for here avoids the mistaken alternative between scepticistic relativism on the one hand and naive realism about knowledge-attribution on the other hand.

A final question: How do we get to our descriptions? What makes us choose this rather than that description? The decision is usually not arbitrary or "subjective" in any of the suspicious senses of this word. It rather depends on what we think is relevant in a situation. And this, again, depends on the practical interests we have in the situation. Think of the clock-example. If I am rather interested in the future (because, e.g., Carol and I need to catch a train), then I tend to ascribe knowledge about the time to Carol. If, however, I am more interested in the past (wanting to reconstruct what happened before Carol looked at the clock and just left), then I tend to not ascribe knowledge about the time to Carol. More generally: When we say about a person that she does or does not know that p, we thereby also express (but do not report) the practical interest which gives us a certain perspective on the situation. But all this does not mean that our way of talking about knowledge is not "objective"

¹² Furthermore (as I said above): False descriptions are excluded from the start.

or not true or false. However, it shows us something different which is not unimportant: that our concept of knowledge has a practical dimension.

REFERENCES

- Foley, Richard, *The Theory of Epistemic Rationality*, Cambridge/MA: Harvard University Press 1987
- Goldman, Alvin I., *Epistemology and Cognition*, Cambridge/MA: Harvard University Press 1986
- Harman, Gilbert, "Knowledge, Inference and Explanation", in: *American Philosophical Quarterly* 5, 1968, 164-173
- Ramsey, Frank Plumpton, "Knowledge", in: Frank Plumpton Ramsey, *Philosophical Papers* (Hg.: D.H. Mellor), Cambridge: Cambridge University Press, 1990, 110-111
- Russell, Bertrand, *Human Knowledge: Its Scope and Limits*, New York: Simon and Schuster 1948
- Unger, Peter, "An Analysis of Factual Knowledge", in: Michael Roth/ Leon Galis (Hg.), *Knowing. Essays in the Analysis of Knowledge*, New York: Random House, 1970, 113-130
- Williams, Bernard, *Descartes. The Project of Pure Enquiry*, Harmondsworth: Penguin 1978
- Zagzebski, Linda, "The Inescapability of Gettier Problems", in: *The Philosophical Quarterly* 44, 1994, 65-73

Peter Baumann
Universität Göttingen
e-mail: pbauman@gwdg.de

REPRESENTACIÓN IMPLÍCITA Y NADA QUE REPRESENTAR: Crítica al *Der Raum* de Carnap

JUAN BAUTISTA BENGOTXEA

Universidad del País Vasco / Euskal Herriko Unibertsitatea

ABSTRACT: The purpose of this paper is to analyse the dissertation of Carnap [*Der Raum*, 1922] with the aim of reconsider it from the point of view of the present-day theories of representation in philosophy of science. The article is separated in three sections: (i) a brief presentation both of the minimal structure of *Der Raum* and of a synthesis about the formal, physical, and intuitive spaces; (ii) an outline of the putative 'representational' analogies between Carnap's (1922) and van Fraassen's (1989) views; and (iii) a proposal of the fundamentals for a critical representational interpretation on spatial knowledge derived from Carnap's own analysis.

KEYWORDS: Space, intuition, structure, representation

El análisis de *Der Raum. Ein Beitrag zur Wissenschaftslehre* [1922] (El espacio. Una contribución a la teoría de la ciencia) (*DR* en adelante), nos permite 'reconsiderar' la obra inicial de Carnap en un doble sentido. Por un lado, *DR* sirve de nexo para comprender en su justa medida las primeras claves filosóficas del pensamiento de Carnap -neokantianismo, fenomenología y convencionalismo, amén de su entronque con influencias siempre explícitas como las de Russell, Frege o Wittgenstein- y, por otro, se reconoce en él un determinado paralelismo general con el modo en que van Fraassen enfoca su concepción del vínculo mundo-teoría. Este examen requiere, no obstante, la clarificación de algunos de los resultados obtenidos por Carnap.

1. El espacio: la intuición en el seno del conocimiento matemático y físico

Con objeto de facilitar una mejor comprensión de las fuentes del conocimiento del espacio, su naturaleza y su ámbito de validez, Carnap dividió su teoría del espacio en tres áreas de una jerarquía determinada por su distanciamiento respecto a lo empírico: (i) en el nivel superior, una

correspondiente al *espacio formal* (R); (ii) en el nivel intermedio, la correspondiente al *espacio intuitivo* (R') y (iii) en la base, una tercera que abarcaba el *espacio físico* (R'').

Los tres espacios, respectivamente, los concibió como ámbitos de estudio del matemático, del filósofo (geómetra) y del físico. Por lo tanto, también las metodologías a aplicar variaban de uno a otro, siendo el *logicismo* la de R, el *convencionalismo* la de R'' y una metodología de tintes *neokantianos* -con reminiscencias epistemológicas *husserlianas*- la del espacio intuitivo (R').

Tras esbozar los rasgos generales de la geometría euclídea, Carnap presenta el espacio formal (DR: 5) no como teoría general de relaciones determinadas (punto, recta, plano), sino en tanto "teoría de relaciones o de orden, es decir, una ciencia de cosas y relaciones completa e igualmente indeterminadas, para las cuales se presuponen ciertos principios a partir de los cuales se deducen otros de forma ilimitada". Como el resto de los espacios, el formal cuenta con tres subcategorías: la *topológica* (t), la *proyectiva* (p) y la *métrica* (m), y sobre cada una de ellas se asignará una variedad dimensional: mono, di, tridimensional, etc. -por ejemplo, espacio formal topológico de tres dimensiones $\equiv R_{3t}$.

La exposición del espacio formal, en la línea de los *Principia Mathematica*, tiene como objetivo básico erigir su propuesta sobre la *teoría de tipos*¹, utilizada con anterioridad para la reducción del número real a la lógica. El caso más general de espacio formal se alcanzaría en las que Carnap denomina "series constantes de nivel superior (series de series)" donde las dimensiones son varias -fundamentalmente tres- y a partir del cual, mediante determinadas particularidades, se obtienen el espacio formal proyectivo (R_p) y "las diversas formas de los espacios (formales) métricos" (DR: 9).

La brevedad del capítulo dedicado a R permite, no obstante, obtener dos conclusiones relevantes. En primer lugar, la crítica de Cassirer contra el logicismo parece no tener cabida en el caso de Carnap. Cassirer arremetía -cf. [Cassirer 1910]- contra *Los principios de la matemática* de Russell (1903), debido a que en la tarea epistemológica de éste se aceptaba que el pensamiento dejara de 'actuar' una vez salía airoso en la *deducción* estricta dada entre sus propias estructuras, abandonando el problema de la *legaliformidad* del mundo de los objetos a la observación directa. La crítica desafiaba a que el logicista diera cuenta del papel que la matemática desempeña en la construcción de la noción de *realidad objetiva*, pues no respondía a la cuestión de la aplicación de las matemáticas a la experiencia. La crítica de Cassirer no será aplicable a

¹ En el capítulo 10, Apéndice B, de [Russell 1903] se desarrolla la primera versión de la teoría de los tipos lógicos. Posteriormente, en los *Principia* de 1910, retoma la cuestión junto a Whitehead, de modo más técnico y sofisticado. Para un esbozo comprensible y correcto, desde la doble perspectiva lógica y filosófica, cf. [Torretti 1998: 2.4].

Carnap, tampoco ya a los *Principia* de 1910, porque en este caso la propia geometría métrica también es reducida a la lógica. Ya no hay peligro de dejar en manos de la experiencia el elemento formal que objetiva, según el neokantiano, la propia experiencia, esto es, las *convenciones métricas*.

En segundo lugar, en vez de acometer el procedimiento hilbertiano de *definición explícita mediante axiomas*, Carnap elabora un estudio de la geometría formal donde los espacios son *objetos lógicos*: espacios definibles únicamente con términos de la lógica; más concretamente, de la teoría de números. Los espacios son ahora objetos *tipo-teóricos* explícitamente definidos en una jerarquía de la teoría de tipos, siendo los axiomas las verdades lógicas acerca de ellos. El espacio formal R, por lo tanto, se ‘convierte’ en una *clase de abstracción* de propiedades y relaciones con aplicabilidad ilimitada. Definitivamente será la componente formal de tales estructuras la que permita que la geometría formal (de R) sea aplicable a cualquier objeto.

El estudio del espacio físico (R'')(DR: 33) comienza con la siguiente interrogante: si es que podemos, ¿cómo hacemos para establecer una línea recta en el espacio físico? No se establecerá únicamente sobre la base de la experiencia, sino que tendrá en cuenta *estipulaciones* libremente elegidas, convenciones sin posibilidad de confirmación o refutación experiencial. Éstas serán de dos tipos, según Carnap:

A/ estipulaciones directas de una clase de objetos físicos (rectos en este caso), o bien

B/ una estipulación métrica que defina una noción de distancia espacial y gracias a la cual la ‘rectitud’ se defina en virtud de *principios geodésicos*².

Carnap se centra en el segundo tipo y comienza distinguiendo entre *forma necesaria* y *forma opcional* (DR: 39):

Denominemos a la materia no amorfa, pero que se da únicamente en la forma necesaria, ‘hecho de la experiencia’. Puede estar sujeta aún a una formación añadida en términos de forma opcional. Para evaluar un enunciado experiencial y saber si es un enunciado de hecho o no lo es, hemos de analizar si el enunciado experiencial es válido para toda formación posible ..., es decir, para todo tipo posible de estructura espacial.

En términos matemáticos, Carnap hace referencia a la *invariancia* –del contenido del enunciado de la experiencia– *bajo transformaciones*, inmutabilidad que sólo se mantiene en el caso de los enunciados topológicos. Tal sería el caso si el contenido de un enunciado experiencial fuera invariante bajo transformaciones (uno-a-uno); esto es, para evaluar un enunciado experiencial y ver si es o no un enunciado de hecho, hay que investigar si es ‘válido’ para todas sus (trans)formaciones –para todos los tipos posibles de

² Para una explicación informal de las *geodésicas*, cf. [Friedman 1983/1991: 62s, 65s]. Para una definición más precisa, *Ibid*: 416ss, 423ss.

estructura espacial-. En este sentido, y aunque eliminativa, la respuesta al motivo por el que las relaciones métricas no se fundamentan en la experiencia, nos remite a la topología del espacio: las propiedades topológicas de éste son las únicas de la noción de *hecho de la experiencia*. La estructura del espacio físico topológico tridimensional (R''_{3t}) se puede erigir sin tener que decidir acerca de la 'rectitud' y de las relaciones métricas de una línea espacial. Las relaciones necesarias para la construcción de estructuras físico-espaciales empíricamente dadas son "las relaciones de incidencia [*betweenness*] de puntos, líneas, superficies y elementos lineales. Sólo R''_{3t} reproduce de modo unívoco lo que está presente en la experiencia" [DR: 37].

Lo importante es descubrir la *forma mínima* de la experiencia que posibilite el conocimiento del espacio físico R'' . Para ello Carnap adopta un enfoque próximo al convencionalismo³ en su vertiente puramente metodológica, no tanto epistemológica, que le permita aclarar la estructura de la física para su posterior aplicación al conocimiento del espacio físico en particular.

Enfoca su tarea a partir de la distinción matemática entre *forma necesaria* y *forma opcional* en los juicios acerca del espacio, lo cual le aproxima más a la *teoría universal invariante del sintético a priori* de Cassirer que a las tesis russellianas de 1914 en *Our Knowledge of the External World*. Entre las conclusiones más interesantes, observamos que hace equivaler forma topológica a forma mínima que un hecho espacial ha de tener para que una experiencia –externa- sea posible, equivalencia que deriva hacia el reconocimiento de que las teorías físicas surgen vía estipulaciones métricas añadidas a la estructura topológica en torno a las cuestiones físicas de hechos espaciales. Precisamente estas estipulaciones serán las que provean las condiciones de matematización de las teorías –caso de la distancia espacial, por ejemplo-. La lógica formal de Carnap apunta hacia la conexión que considera existente entre la forma necesaria de la experiencia espacial y el *sintético a priori*:

Según Kant, el espacio es la condición de posibilidad de toda experiencia (externa) como tal. ¿Es esto verdad para las determinaciones espaciales de las estructuras que hemos distinguido? Para decidirlo debemos considerar qué determinaciones espaciales se dan necesariamente en toda experiencia, y cuándo esa

³ En referencia, principalmente, al convencionalismo de Henri Poincaré. En torno a las cuestiones del espacio, cf. particularmente [Poincaré 1902/1963: 2ª Parte, apartados III, IV y V]. Para una visión global de las tesis convencionalistas y las sucesivas críticas empiristas, vid. [Sklar 1974: Cap. II, F]. Michael Friedman [1991/1983] presenta un estudio más actualizado subrayando los problemas de infradeterminación, relativización, congruencia e inferencia teórica [Cap. 7]. También es filosóficamente interesante la síntesis de [Nerlich 1994: Caps. 6 y 7].

experiencia aún no se ha dado en la forma espacial que va más allá de la forma necesaria en virtud de determinaciones libremente elegidas. Ahora bien, hemos denominado experiencia a la cuestión de hecho [hecho] –en tanto está presente sólo en esa forma necesaria unívoca que no contiene cualquier estipulación arbitraria. Por lo tanto, únicamente las determinaciones espaciales contenidas en los hechos pueden ser las condiciones de posibilidad de la experiencia. Y éstas, como hemos visto, son sólo las relaciones topológicas, no las proyectivas y, menos todavía, las relaciones métricas [DR: 65].

Con esto Carnap adopta el *sintético a priori* en su vertiente *universal invariante*⁴: el estatus objetivo del hecho viene a residir en su sujeción a la *forma necesaria*. Ya no hay necesidad de aludir, con Kant, a la estructura objetivante de la mente, pues Carnap no explica tanto como elucida la forma lógica y topológica.

El objetivo de Carnap es, por tanto, proveer contenido *a priori* a la geometría, de modo que le conceda un carácter *propriadamente espacial*. El medio, instrumento, para tal provisión se sustenta en la intuición. Carnap usa ésta para garantizar de modo *apriorístico* que los hechos topológicos de la experiencia se puedan ordenar en una estructura métrica global unívoca; esto es, presenta la intuición como garante de la metrizableidad del espacio físico, como intermediario no eliminable en el entendimiento de los espacios formal y físico.

Ahora bien, ¿cuál es la noción de intuición que Carnap utiliza? Es precisamente aquella que rescata de la *idea esencial* de Husserl. Según éste, en la idea esencial pueden comprenderse inmediatamente las características universales que los particulares ejemplifican –el color general, la figura espacial general, el tono general-. Por lo tanto, este tipo de idea nos pone en contacto con características universales, y no tanto con individuos particulares; es independiente de los individuos particulares que existen actualmente en un ámbito empírico, pues es fuente *a priori*, no empírica. Mediante la idea esencial se puede conocer *a priori* la estructura del espacio del color, del espacio tonal o, en geometría pura, del espacio geométrico puro. Sin embargo, este conocimiento no es analítico, sino sintético; no se sigue solamente de las verdades generales de la lógica formal pura –que expresa la *esencia* [Eidos] común a todo objeto del pensamiento-; más bien sirve para subespecies particulares de objetos. Esto se debe a que la intuición tiene una componente *propriadamente espacial*, diferente de la de otros elementos fenoménicos también aprehensibles mediante la idea esencial. Hay un modo *a priori* ‘primario’ de aprehensión que distingue las siguientes características:

⁴ En sintonía con Cassirer y su *teoría universal invariante de la experiencia*, donde lo formal del conocimiento empírico es la *forma mínima* que la experiencia ha de tener para garantizar la posibilidad de las convenciones.

características espaciales ≠ características de color, audición u otras de la intuición de objetos

La esencia de estas características espaciales permite relacionar, por tanto:

estructuras lógicas formales de la geometría matemática @ experiencia de relaciones espaciales

Una vez dadas la distinción y la relación indicadas, es importante tratar de descubrir con la intuición espacial lo que realmente se pueda. Resulta obvio que se obtiene un conocimiento de magnitud limitada; la intuición tiene que ver únicamente con una región espacial limitada. Y, al igual que sucede con las series en el espacio formal, estas formas espaciales pueden iterarse mediante un proceso de adición (DR: 23). En resumen, la intuición se aplica a regiones ‘pequeñas’ y permite la iteración. Ambas conclusiones dan cabida al estudio carnapiano del *espacio intuitivo* (R'), que desarrolla en dos fases: en la primera, mediante los axiomas de Hilbert, pretende encontrar cuáles de éstos provienen de la intuición de una región. En la siguiente, investiga qué postulados –consistentes– se pueden añadir a los axiomas para así determinar una estructura global para R' .

La resultante de tal tarea es una estructura o clase general de *variedades riemannianas* metrizable tridimensionales. Por lo tanto, se consigue una generalización de la noción de variedad métrica a una clase de variedades tridimensionales que permiten una métrica, aunque no tengan definida una sobre todas ellas. Carnap apoda a esta conexión con la métrica, *espacio intuitivo métrico tridimensional* (R'_{3m}) –aun siendo consciente de que no conlleva métrica determinada ninguna.

Ahora bien, no toda espacio topológico es metrizable; incluso si los hechos experienciales aportaran conocimiento fiable y completo de la estructura topológica del espacio físico, tampoco tendríamos garantía de poder construir una convención métrica de modo unívoco. Es la estructura del sistema intuitivo la que viene a ofrecer, según Carnap, una garantía *a priori* para la creación de tal métrica, ya que la estructura de la intuición es ‘impuesta’ sobre el espacio físico. Lo que convierte al espacio físico en ‘espacio’ genuino es su relación con la intuición propiamente espacial – a diferencia de, por ejemplo, su relación con la intuición auditiva, pues la variedad tonal auditiva no es espacial-. Y en este contexto son las convenciones métricas las que desempeñan el papel ‘sistematizador’, aunque siempre a costa de añadir a los hechos de la experiencia la estructura *a priori* del espacio intuitivo.

Por último, Carnap usará el instrumental del espacio formal para generalizar el conocimiento dado en el espacio intuitivo a estructuras más abarcales, hasta llegar a lo que denomina *espacio intuitivo topológico de n dimensiones* (R'_{nt}), el más abarcaliente, continente de todos los espacios intuitivos posibles (DR: 31); en suma, *la condición de posibilidad de cualquier objeto de la experiencia*.

2. De la intuición a la construcción conceptual

Como se ha indicado, en el momento de determinar cuál es el espacio de nuestro universo físico, surgía la dificultad subyacente al enfoque kantiano del carácter *a priori* de la geometría euclídea como geometría del espacio de la intuición pura. La geometría euclídea no es *a priori* porque no está unívocamente determinada como forma del espacio físico. Según Carnap, el espacio que sí muestra ese tipo de 'determinatividad' –de estatus *apriorístico*– es el espacio topológico. Y ello es debido a que queremos dividir lo que nos es dado integralmente en la experiencia en dos componentes formales :1/ la forma necesaria y 2/ la forma opcional.

Este enfoque basado en la distinción entre *forma necesaria* y *forma opcional* de los juicios espaciales entronca con las formas topológicas o formas mínimas del espacio. Por lo tanto, la forma necesaria es la forma mínima que un hecho espacial debe tener para que la experiencia sea posible –el candidato a *forma a priori de la intuición pura*, fundamentado principalmente en la *teoría universal invariante* de Cassirer–.

Sin embargo, la relación entre espacio físico y espacio intuitivo, que Carnap denomina relación de *subsunción* o *subordinación* [*Unterordnung*], presenta ciertas dificultades que giran en torno a su supuesto estatus representacional (*representación implícita*). Esta relación de subsunción de fenómenos físicos experimentados a la forma sintética *a priori* del espacio intuitivo, concede al espacio físico el carácter de dispositivo ordenador de los objetos de la experiencia en el espacio intuitivo construido *a priori*. Recordemos que el sentido de ello residía en que la *idea esencial* surge, en un primer momento, de nuestra experiencia sensible de los objetos espaciales. Tal experiencia –espacial– nos permite llegar a las leyes *a priori* que gobiernan la estructura del espacio intuitivo. El espacio físico, por tanto, resulta ser un ordenamiento preciso y consistente de objetos espaciales en una estructura matemática, siendo la tarea de ésta asignar relaciones espaciales a objetos –sea la determinación de ángulos y distancias mutuas, por ejemplo– tal que se puedan formular leyes matemáticas exactas.

El problema radica en que el espacio intuitivo, tal y como lo entiende Carnap, posee sólo estructura topológica. Más precisamente, posee únicamente una estructura infinitésimamente euclídea, de modo que sólo *a priori* se nos es dada la clase de todas las variedades riemannianas posibles. Ahora bien, hemos de tener en cuenta que el espacio físico requiere otro tipo de estructura matemática, la métrica completa –local y global–, lo cual obliga a seleccionar una variedad riemanniana particular. Esto se consigue, según Carnap, incorporando tal métrica mediante una convención estipulada libremente. Alternativamente, se puede partir de una estipulación de la medida.

Lo importante es que la estructura métrica a la que se llega no es inherente a los hechos, sino que tan sólo expresa el resultado de nuestra libre elección –no esencial-. ¿A qué se debe, pues, esta independencia de la estructura métrica respecto a los hechos? En la respuesta de Carnap entran en juego las nociones de *cuestión de hecho* [hecho] de la naturaleza [*Tatbestand*], *forma necesaria* y relaciones topológicas.

Si Carnap concebía la geometría como ciencia *eidética* factualmente dada cuya esencia se nos revelaba en actos del análisis de la esencia inmediatamente intuitivos, casi perceptuales, nosotros rechazamos tal modo de representación y proponemos en el que el conocimiento geométrico no se nos es dado directamente a la intuición, sino que expresa una construcción –representacional- con base anti-intuitiva e instrumental matemático. Y aunque nuestra inteligibilidad físico-matemática del espacio pueda comenzar con actos intuitivos, el resultado de un procedimiento así, lo reiteramos, es anti-intuitivo, puramente conceptual, una construcción simbólica mediante la que se representa el mundo espacial y físico gracias a símbolos matemáticos abstractos sin contenido intuitivo. Sólo gracias a tal construcción simbólica se puede obtener un conocimiento objetivo del mundo físico:

Esta concepción de la objetivación se aproxima más al neokantismo de Cassirer que a los rasgos fenoménicos de Carnap o Husserl. El espacio de la teoría de la relatividad, por ejemplo, trasciende nuestra intuición espacial; es ya una estructura puramente abstracta, conceptual, cuya naturaleza viene expresada en la circunstancia de que el espacio tangente en cada punto de la variedad matemática es euclídeo. Y lo que determina empíricamente la orientación mutua contingente de tales espacios tangentes es, precisamente, la ecuación de campo einsteiniana, pues sólo la teoría de Einstein ofrece una construcción simbólica adecuada de la realidad física. Carnap, por el contrario, concebía la naturaleza del espacio euclídeo *a priori* como *reflejo –encaje* si hubiera aplicación- de la estructura necesaria (forma topológica) de nuestra intuición espacial –adelantándose en cierto modo a un tipo de representación *correspondencialista* especular-.

3. Hacia una representación sin encajes

Las dificultades de admitir un espacio, el intuitivo, en su especie topológica como substrato objetivo para la construcción del espacio físico, chocan en cierta medida con la noción de objetividad proveniente de las invariancias de una concepción relacional como la de Cassirer. El sentido carnapiano de la objetividad es inverso y de tintes casi fundamentalistas. No se dirige hacia la formación de un concepto dinámico de objetividad, basado en la no variabilidad que a lo largo del proceso representacional del hecho surge a flote. Más bien busca el fundamento primero, absoluto, del espacio para que los hechos y sus objetos tengan en él cabida.

Una alternativa más convincente es la de no aislar el conocimiento del hecho mismo, del objeto espacial que con sus propiedades induce a conseguir aquella esencia a la que Carnap se refería, vía intuición esencial. Es decir, partir del hecho experiencial –incluyendo su ‘lo dado’, es decir, sus objetos y propiedades (espaciales)- manteniendo la tesis de que éste es el *inicio* del conocimiento. Esta perspectiva, antecedente cassireriano del representacionalismo formal actual en filosofía de la ciencia, muestra al objeto como producto de un proceso de constitución cognoscitiva. Tal y como bien relatan [Ibarra/Mormann 1997: 251], “los objetos no se encuentran, por tanto, más allá de la teoría, sino que se constituyen mediante la teoría, esto es, mediante leyes conceptuales”. No hay necesidad de manejar nada *a priori* en lo factual; lo que se requiere es un proceso dinámico que lo constituya. Cassirer (1910) caracterizó funcionalmente la formación de los conceptos en el contexto del desarrollo de una serie determinada por ciertas invariantes, versión próxima en su programa al de Klein y que, en su vertiente empírica, remite a la *gradualidad* a la hora de responder a si nuestro conocimiento es objetivo o no: lo es más o menos. Esto es, no hablamos de *cosas en sí* vs. *representaciones*; ya no decimos que lo objetivo se dé como algo fijo que está ahí, como *forma necesaria*. “La objetividad no es un concepto estático: la realidad o, mejor aún, la realidad científica [espacial], según Cassirer, no está *ahí* simplemente, sino que se constituye en una sucesión de sistemas conceptuales relacionales progresivamente más complejos y coherentes” [Ibarra/Mormann 1997: 260s].

Si la estructura matemática de una teoría científica puede, en diversos enfoques semanticistas de filosofía de la ciencia, obtenerse mediante una *definición teórica* –caso de [van Fraassen 1989]-, la vinculación de la teoría, de aquellos modelos matemáticos, con los sistemas reales –*hecho experiencial* en Carnap- requerirá por su parte lo que van Fraassen denomina la *hipótesis teórica*, según la cual, una determinada clase de sistemas reales (M_R) está contenida en la clase de las estructuras (modelos M_T) que la definición teórica identifica. La alusión al enfoque de van Fraassen⁵ surge del paralelismo con la propuesta de relación entre espacios de Carnap. Éste, muy sucintamente, propuso dos interpretaciones:

1/ el espacio formal y sus términos primitivos no interpretados reciben una interpretación específicamente espacial en el espacio intuitivo: relación de *especificación* [*Einsetzung*]

2/ pero la interpretación de carácter intuitivo del espacio formal no equivale a una interpretación física. Ésta se obtiene mediante la relación de *subsunción* [*Unterordnung*]

⁵ Para una introducción de la concepción semántica de este autor, véase [van Fraassen 1989: III.9].

Y, al igual que en [van Fraassen 1989], donde la identificación de los modelos de la teoría –mediante el *espacio de estados*- se construye a partir de las nociones de *transformación*, *invariancia* [y *simetría*], Carnap también fundamentaba su constitución del espacio intuitivo (R') sobre la teoría universal de la invariancia. Y es el espacio intuitivo, precisamente, el que en su relación con el formal, desempeña el papel de estructura global donde se incluyen subestructuras [M_R o espacio físico] (*representación implícita*, diríamos) al modo no sofisticado de $M_R \subset M_T$.

No obstante, en sus interpretaciones ‘empíricas’ (física), tanto van Fraassen como Carnap rehusan desarrollar un concepto funcional, más consistente que el de inclusión directa, con un enfoque invariante. Encontramos una propuesta más interesante en este sentido en el seno de los enfoques representacionistas de filosofía de la ciencia⁶. Sin entrar en detalles de reconstrucción, señalaremos tres características que consideramos reflejan más adecuadamente la relación carnapiana con el *hecho* experiencial espacial:

A/ en primer lugar, la perspectiva representacionista asume tanto el papel constructivo del proceso de representación como las características relacionales de los sistemas representante y representado

B/ en segundo lugar, tiene en cuenta la relación funcional entre sistemas (teorema de representación)

C/ por último, el teorema de unicidad inherente a la representación refleja las características invariantes en el proceso constructivo

De modo general, las *fases* para la construcción de una representación consistirían en la conceptualización de los siguientes ítems:

1ª/ Fase de los *datos*: las invariancias o ciertas relaciones permiten obtener los datos. En el caso de Carnap, la intuición en un primer momento, y la invariancia en las secuencias de hechos experienciales permitirían la consecución de tales datos –y no tanto un *a priori* inamovible y objetivo-.

2ª/ Fase *fenoménica*: donde los datos –de la intuición, en Carnap- reciben una representación formal.

3ª/ Fase *relacional-funcional*: se construye el entramado relacional-funcional en cada ámbito. En el caso de Carnap, en los dominios intuitivo-matemático y en el físico, teniendo en cuenta las *leyes* de cada uno.

4ª/ Fase de *síntesis*: consecución de la representación y sus valores epistémicos añadidos: descripción, explicación, objetivación.

BIBLIOGRAFÍA

- Carnap, Rudolf: 1922, “Der Raum. Ein Beitrag zur Wissenschaftslehre”, *Kant-Studien*, Ergänzungshefte 56, Berlin.
 Cassirer, Ernst: 1910, *Substance and function*, New York, Dover, 1953.

⁶ Cf. las siguientes referencias: [Mundy 1986], [Swoyer 1991], [Ibarra/Mormann 1997: 3.1, 4.1].

- Cassirer, Ernst: 1921, *On Einstein's Theory of Relativity*, New York, Dover, 1953.
- Friedman, Michael: 1983, *Fundamentos de las teorías del espacio-tiempo. Física relativista y filosofía de la ciencia*, Madrid, Alianza, 1991.
- Giere, Ronald N., Richardson, Alan W. (eds.): 1996, *Origins of Logical Empiricism, Vol. XVI, Minnesota Studies in the Philosophy of Science*, Minneapolis, University of Minnesota Press.
- Howard, Don: "Relativity, Eindeutigkeit, and Monomorphism: Rudolf Carnap and the Development of the Categoricity Concept in Formal Semantics". En R.N. Giere & A.W. Richardson (eds.): 1996: 115-164.
- Ibarra, Andoni, Mormann, Thomas: 1997, *Representaciones en la ciencia. De la invariancia estructural a la significatividad pragmática*, Barcelona, Ediciones Del Bronce.
- Mundy, Brent: 1986, "On the General Theory of Meaningful Representation", *Synthese* 67: 391-437.
- Nerlich, Graham: 1994, *The shape of space*, Cambridge, Cambridge University Press.
- Poincaré, Henri: 1902, *La ciencia y la hipótesis*, Madrid, Espasa-Calpe, 1963.
- Richardson, Alan W.: 1998, *Carnap's construction of the world. The Aufbau and the emergence of logical empiricism*, Cambridge, Cambridge University Press.
- Russell, Bertrand: 1903, *Los principios de la matemática*, Madrid, Espasa-Calpe, 1977.
- Sklar, Lawrence: 1974, *Space, Time, and Spacetime*, Berkeley, University of California Press.
- Swoyer, Chris: 1991, "Structural Representation and Surrogate Reasoning", *Synthese* 87: 449-508.
- Torretti, Roberto: 1998, *El paraíso de Cantor. La tradición conjuntista en la filosofía matemática*, Santiago de Chile, Editorial Universitaria y Universidad Nacional Andrés Bello.
- van Fraassen, Bas C.: 1980, *La imagen científica*, México DF, UNAM-Paidós, 1996.
- van Fraassen, Bas C.: 1989, *Laws and Symmetry*, Oxford, Clarendon Press.

Juan Bautista Bengoetxea
UPV/EHU
e-mail: ylbbecoj@sc.ehu.es



LIMITS IN SCIENCE

DRAGANA BOZIN
University of Oslo

ABSTRACT: Usually, debates concerning limits of science gravitate towards showing the existence of specific limits. The term itself appears in various contexts that range from political and moral to metaphysical, frequently in one and the same text and with no clear demarcation lines. In order to make any serious advances in research on the nature and the role of limits in science, it is necessary to consolidate the effort. This cannot be done successfully without definitions. I present an example of the confusion that results from employing the word "limit" to mean different things and some suggestions as for future research.

KEYWORDS: Limits, science, knowledge.

Introduction

Articles on scientific limits generally focus on showing the existence (or non-existence as it might be) of a specific limit. Often these limits are very different in character.

Take for example¹ the uncertainty principle in quantum mechanics. One way to describe this principle is to say that it is not possible to predict or measure both the position and the momentum of a particle at one time. An alternative (and more accurate) interpretation is that there is no state of a quantum mechanical particle with a precisely defined position and momentum and that the uncertainty principle, therefore, is not a limit to what observed properties of a quantum particle are predicted by the theory. Since there is no quantum state with precisely defined position and momentum, quantum theory predicts that we shall never observe both simultaneously.

Another example² is provided from cosmology and concerns the actual visual horizon. Observers at any time of the history of the universe can only see galaxies within a distance close enough that their light could have reached them in the time since the big bang. The final velocity of light ensures that we

¹ The example is taken from John L. Casti, 1996

² John L. Casti, 1996

have a visual horizon (about 15 billion years away) in the universe, beyond which the light has not had time to reach us since the expansion of the universe began. What can be studied is a finite portion of the universe which might not contain enough information to determine the nature of initial state for the whole universe without the introduction of unverifiable, and necessarily highly speculative 'Principles' to which the initial state is believed to adhere.

We encounter still more different kinds of limits like for example the intrinsic limits to what can be predicted and verified in science³. Or even more different ones like the limits imposed on science by ethical considerations (gene-technology, experimentation on humans etc.), political constraints, current social interests and so on.

An attempt to understand the concept itself or to map out various kinds of limits in science escalates quickly, revealing an entire maze of possible meanings and interpretations. In addition the question what science can or cannot answer merges easily (and understandably) with closely connected and not readily answerable question of how much we can know or understand in principle.

This situation does not lend itself to more systematic investigation into the nature as well as the role of limits in science. In the following section I will show how multiple concepts/meanings figure in an example text on scientific limits. The two subsequent sections contain a short discussion of these examples and some suggestions on how to proceed to consolidate investigations of limits in science.

Examples

I take as a point of departure (rather than a critical assessment of) an article by James Hartle in "Boundaries and Barriers: on the Limits of scientific knowledge"

The title of this collection of essays is a good representative of the literature in this field: on the one hand it seems to indicate that these words (limits, boundaries and barriers) have different meanings. No clear distinction or even an attempt to make one is advanced in the book. No clear statement (or any statement at all) was made to suggest that these words do have the same or at least sufficiently similar meanings.

In his article Hartle makes the following two main points concerning limits:

- In general limits of scientific knowledge concern the relationship between observation and theory. Limits depend on theories and will vary from one scientific theory to another

³ This particular limit is discussed below

- There are (at least) three different kinds of limits, general in character and inherent in the nature of the scientific enterprise: limit to what is (can be) predicted, limit to implementation and limit to what is (can be) verified.

Specific limits are further elaborated as follows: theories do not predict everything that is observed but only certain regularities in our observations. Scientific laws must have some degree of simplicity (to be discoverable, comprehensible and applicable by human beings) The complexity of the universe and the simplicity of laws would imply a limit.

To be tested, predictions must be implemented in particular circumstances. This often involves computation. Even if the laws are precisely specified and even if the input to those laws is exactly stated, limitations of our ability to compute may limit our ability to predict. In addition there are limits to what a machine can do; certain predictions may be non-computable numbers (the precision required of initial data to extrapolate a given time into the future increases exponentially with that time for a wide variety of classical systems).

And finally limits to verification: the laws of physics are a result of induction and experiment and consequently never 'known' to be true.

Discussion

The three kinds of limits Hartle mentions are described in very loose terms, using statements of varying rigor and clustering several kinds of limits under the same heading. For example: there are two statements explaining the limit to prediction (1) Out of the infinite number of observations, we predict (presumably cover by laws) only certain regularities and (2) If the universe is complex and the laws simple then there must be a limit to scientific knowledge.

We could say that the statement (1) employs a rather 'weak' sense of the word limit - our laws indeed cover a limited number of regularities we observe (the rest might or might not be of vital interest to us) but that number changes as do our theories. Although it might be the case that science will never predict all the regularities, this kind of limit is a moving target. The statement (2) makes a much stronger metaphysical claim: there is a limit that results from the nature of the physical world and the nature of scientific laws. This limit does not change from theory to theory (it holds for all theories) and no amount of scientific knowledge could lessen it. In addition to these two statements the same paragraph in the text slips into a third kind of limit (laws must be simple to be discoverable, comprehensible and applicable by human beings) which is a limit of our ability to understand.

Limit to what can be implemented creates a similar kind of confusion among different kinds of limits. There is again an easy transition from our ability to compute, to practical failure to compute, to non-computability (in principle).

Limit to what can be verified seems to be yet another specimen: limit as a consequence of the problem of induction.

Without elaborating the points above I would like to point out that the degree of imprecision in the use of this crucial term makes it impossible to assess the arguments that follow (in the article) Furthermore they create a kind of ceiling on discussions on limits in science.

Suggestions for further development

More rigorous stipulation of the meaning(s) of the word limit and/or classification would be necessary to make a more systematic investigation into the role limits have in science.

One way is to follow areas of philosophy: to differentiate for example between (at least) physical, metaphysical, epistemological, cognitive, and ethical limits and then work on defining and explicating the role in science for each one in turn.

Another possibility could be to focus on defining the concept of limit. (These two options can of course be combined.) In fact it is likely that more than one concept will be needed to cover so many different examples of limits in science. Let us start then with the following three: limit, boundary and barrier.

Let us say further that science 'encounters' a barrier in those cases where it cannot provide an explanation (prediction, measurement, etc.) for reasons that are temporary and 'removable' as it is for example the case with limitations that science faces in the development of technology, regarding ethical dilemmas, politic, etc.

The concept of limit can be defined by 'mimicking' the respective definition in mathematics. In very loose terms, the limit of a function or a sequence is a number that function/sequence can approach infinitely or alternatively 'reach' it (for function – to has that number as it's value, for sequence – to be a member of it). A function can never have a value greater than its limit; a sequence can never have a member greater than its limit. Again speaking loosely the limit is permanent, a kind of 'final frontier'.

Let us define limits in science as those cases where science cannot provide an explanation (prediction, measurement, etc.) for reasons that are not possible to change or 'remove'; 'points' that scientific knowledge can 'reach' but never 'pass'. If we now apply this definition to Hartle's text (2) qualifies as a limit and (1) does not.

Boundary could be defined as a demarcation line between areas, as in for example boundaries between sciences. This interpretation would not be completely unlike the related concept in mathematics where for example a sequence is bounded by all the numbers greater or lower than any of it's members. The concept of boundary could be a kind of bridging concept between the concept of limit and the concept of barrier in the sense that it

could be changed or 'moved' but not 'removed'. On this interpretation (1) from Hartle's article would be a boundary and (2) would not.

REFERENCES

Casti, John L. and Karlquist, Anders: *Boundaries and Barriers: on the limits to scientific knowledge*, Addison-Wesley Publishing Company, Inc. 1996.

Dragana Bozin
University of Oslo
e-mail: dbozin@uio.filosofi.no





LA INTERPRETACIÓN DE LAS ESTRUCTURAS SOCIALES (DE LOS OTROS) Y EL INDIVIDUALISMO METODOLÓGICO

FERNANDO BRONCANO
Universidad de Salamanca

ABSTRACT: Social structures inherit a sort of opacity from the representations of individuals that interact inside them. Because it, a direct explanation is not possible and we need to identify and interpret what are the structures of a community. The claim of interpretation poses a similar problem for methodological holism and individualism. The paper aims to extend the Charity Principle from the individual to social contexts. We ground this extension in the social abilities created by the *Theory of Mind* mechanism. Social expectations about others allow to explain an Equilibrium Principle in each social intercourse. And, beyond this equilibrium, systematic biases are viewed as social patterns that constitute the social structures

KEYWORDS: Philosophy of social sciences, action theory, charity principle, interpretation, equilibrium.

1. La controversia sobre el individualismo metodológico y la transparencia intercultural de las instituciones

Es sabido que la tensión entre individualismo y holismo nace de las dificultades de explicación que presentan las estructuras genuinamente sociales, que demuestran ser irreductibles a conjuntos de conductas individuales. El teórico social individualista está obligado a mantener la autonomía del nivel institucional ya que las instituciones se convierten, una vez creadas, en constricciones casi físicas para la acción individual. El holista, por su parte, tiene que explicar cómo es posible esta autonomía sin postular algo así como una mente colectiva. ¿Cómo surgen y cómo se mantienen las instituciones sin presuponer otras instituciones anteriores?

El hecho sorprendente es que ninguna de las dos concepciones parece encontrar excesivos problemas epistemológicos con las estructuras sociales: ambas consideran que las estructuras sociales son epistémicamente *transparentes*. A diferencia de lo que ocurre con las intenciones subjetivas de

los agentes, las estructuras sociales implicadas en las acciones - parece inferirse del silencio acerca de este problema - se pueden ver directamente o deducir de la conducta observada. Pero no hay nada que justifique este optimismo; las relaciones e instituciones sociales no son más transparentes que las intenciones. Y en este terreno el individualista metodológico se encuentra en mejor posición que el holista para dar cuenta de este fenómeno, puesto que sabe que las instituciones sociales heredan de las intenciones individuales el problema de la opacidad.

La opacidad de las intenciones, es bien sabido, depende de algunas propiedades que tienen las representaciones psicológicas y que no tienen otros fenómenos empíricos: en primer lugar, las representaciones tienen esencialmente un *punto de vista* o *modo de presentación* que es subjetivo, a diferencia de los conceptos, que son públicos. En segundo lugar, son independientes de las condiciones causales externas: una representación tiene su propia dinámica ajena a la realidad, de manera que observando la conducta externa del individuo no podemos hacer inferencias directas sobre sus representaciones. Estas dos propiedades son fundamentales para nuestra argumentación, pues las instituciones sociales las heredan de los estados representacionales. Al igual que las representaciones son diferentes en cada individuo, las instituciones y relaciones sociales se realizan de forma distinta en diferentes sociedades. Por eso tenemos dificultades para decidir si las semillas de cacao son o no dinero en la cultura de los incas. Quizá se pueda argüir que se trata de una cuestión nominal, que siempre es posible definir clara y distintamente las condiciones de una institución para que funcione como moneda: es suficiente con que funcione como sistema universal de intercambio. Pero estaríamos cometiendo un grave error de transposición desde *nuestra* institución de dinero. Podemos establecer como condiciones necesarias y suficientes las normas que establecen nuestros sistemas monetarios, está bien, pero vamos a tener que redefinir continuamente lo que es el dinero. Porque al final el dinero es un patrón de disposiciones colectivas a intercambiar bienes bajo ciertas normas y estas disposiciones se realizan en diversos sistemas de intercambio pero no se agotan en ellos. De modo que establecemos normas positivas que “definen” una institución en un momento dado, pero sabemos muy bien que no la constituyen. Pensemos como contraejemplo en los sistemas de regalos mutuos que son empleados en muchas culturas como medios de intercambio que establecen obligaciones mutuas calculables. O, en el extremo contrario, en una cultura en la que las transacciones a crédito, a través de tarjetas, por ejemplo, hayan acabado definitivamente con el soporte moneda del dinero.

En segundo lugar, las instituciones y relaciones sociales son independientes de los procesos causales que las instituyen. Precisamente esta independencia es el pecado capital del funcionalismo como explicación de las instituciones: las normas por las que se sostiene una institución como la

familia biparental y monogámica son independientes de las causas selectivas por las que se impuso.

En el momento en que nos situamos en un contexto de traducción radical descubrimos el hecho de la opacidad de las instituciones sociales y es entonces cuando se muestran más claramente las dificultades metodológicas del holista metodológico puro: un holista puro no tendrá inconveniente en reconocer que otras culturas están formadas por instituciones diferentes, y no obstante, en contextos de interpretación radical, será incapaz de explicar cómo es posible comprender las conductas de esa cultura.

El holista metodológico cree que las instituciones sociales son transparentes precisamente porque son independientes de la fuente de opacidad que es la mente subjetiva de los agentes individuales. Reconocer un cierto grado de opacidad en las relaciones sociales es casi como reconocer que las relaciones son fruto de algo así como una “mente” colectiva. El individualista, por el contrario, está en disposición de explicar cómo la opacidad de las representaciones se hereda al constituir instituciones sociales, pero tendrá que pagar un precio que probablemente socave la forma más tradicional de individualismo: tendrá que abandonar la equivalencia entre individualismo metodológico y atomismo metodológico.

Hay que diferenciar entre los aspectos ontológicos del individualismo y sus compromisos con un determinado modelo de persona: el individuo aislado de las influencias ajenas, perfectamente racional, perfectamente autointeresado. El individualismo sostiene la posición ontológica de que todas las estructuras sociales reciben sus capacidades causales de las disposiciones de conducta de individuos. Los individuos son agentes intencionales y en consecuencia toda explicación debe hacer referencia en último extremo a las motivaciones y creencias de los agentes. En palabras de P. Pettit, 1993 el individualismo es la tesis ontológica de que *no existen mentes colectivas*. El atomismo, por el contrario, añade algo más a esta tesis individualista: es una tesis sobre la constitución de las capacidades sociales de los individuos. Consiste básicamente en una tesis internalista acerca del contenido de los estados mentales sumada al individualismo. El contenido mental no exige referencia a objetos externos y se agota en las relaciones lógicas internas que construyen sus componentes conceptuales y, en consecuencia, en el contexto de las relaciones sociales los otros agentes son para el individuo meras representaciones internas. Pero los agentes humanos son, en los contextos de las relaciones sociales, intérpretes continuos de la mente de los otros a través de la interpretación de su conducta. Las conductas de los otros se convierten en el medio con el que se instauran relaciones que, desde una concepción externalista del contenido mental, forman parte de los propios contenidos.

En la concepción individualista tradicional, el individuo está doblemente aislado. En primer lugar por el postulado internalismo de las representaciones, autónomas respecto al discurrir causal de la realidad, en

segundo lugar por la indiferencia a los intereses de los otros. En el extremo contrario al internalismo encontramos una forma radical de externalismo que ha sido denominada como “la mente extendida” (Clark, 1998, D. Chalmers y A. Clark, 1998): los propios objetos del mundo, por cuanto son necesarios para sostener la actividad mental, se convierten en auténticos instrumentos del cerebro. En contextos sociales (ésta es la enseñanza de la aproximación de la mente extendida) son los otros quienes se convierten de una cierta manera en un elemento necesario de la propia actividad mental. Son, como veremos inmediatamente, auténticos instrumentos de la propia mente.

2. Las capacidades de socialización y las relaciones sociales

Una promisorio línea psicológica reciente considera que la capacidad de relación social es una capacidad muy ligada a la estructura de la mente de nuestra especie, (quizá con algunas habilidades precursoras en las mentes de los primates más cercanos). Se ha denominado a esta capacidad *Teoría de la Mente* y, aunque el término es discutido y discutible, ha conseguido cierta popularidad por lo que nos referiremos a ella con este nombre. Esta capacidad sería innata y tendría su propia estructura cerebral en la forma de un módulo con procesos de desarrollos bien determinados. Los casos de autismo indicarían defectos en este módulo con las consecuencias conductuales que unimos al síndrome del autismo: incapacidad de interpretar la mente de los otros y, en consecuencia, incapacidad radical para relaciones sociales genuinas.

Para unos se trata de un mecanismo de filtro teórico de las conductas: el cerebro humano tendría una especie de “teoría de la mente” análoga a la “teoría euclidiana del espacio”. Mediante esa teoría las conductas de los otros serían clasificadas, construidas y predichas a través del instrumento básico de la psicología natural o “folk” de las creencias/deseos/acciones. Para otros no existe tan constructo teórico sino una capacidad de simular la mente de los otros en la mente propia y un mecanismo de producción de una predicción acerca del otro. Sea cual sea la estructura correcta de interpretación, se trata de un mecanismo esencialmente dirigido a la práctica de las interacciones sociales. Desde la perspectiva del agente, pongamos por caso un niño en las primeras etapas de socialización, la conducta de los otros contiene las suficientes claves como para poder predecir la conducta del otro y, en su caso, ajustar la propia al plan de interacción deseado. Si los otros fueran seres no mentales y no creativos, la predicción solamente exigiría una mera sintonía más con regularidades de conducta más o menos compleja. Un ser conductista dotado de una forma de aprendizaje por refuerzo sería capaz de sobrevivir en un contexto social. Pero los contextos sociales humanos contienen un repertorio ilimitado de pautas de conducta que solamente pueden ser predecibles mediante el recurso a las motivaciones y al estado mental del

agente. La capacidad de socialización exige una sintonía con “la mente” o el contenido mental del otro a través de su conducta.

Es muy probable que la capacidad de socialización haya evolucionado en el contexto de los grupos altamente estructurados por relaciones sociales de dominación, jerarquías, parentesco estable e incluso afinidad (Whiten, y Byrne 1988 y 1997). En estos contextos la capacidad de leer la mente del otro a través de las claves de su conducta tiene claramente una ventaja competitiva. En estos contextos no es difícil encontrar casos de coaliciones estables o esporádicas para conseguir ascensos en la jerarquía del grupo, engaños, pero también cooperaciones y mecanismos de mantenimiento de la estructura del grupo por encima de las presiones centrífugas de las competitividades individuales. El salto evolutivo de monos a los grandes simios consiste en la sensibilidad a los estados mentales de los otros. Probablemente el salto evolutivo a los homínidos consistió en la capacidad completa de “modelizar” la mente ajena a través de una psicología natural de creencias/motivaciones.

De modo que volvemos al problema de la no transparencia de las instituciones sociales. Nuestras capacidades de socialización deben tener una capacidad correlativa de interpretar las relaciones y patrones estables fruto de estas capacidades. Insisto de nuevo en que esa capacidad de interpretación debe ir más allá de la interpretación de la conducta individual: debe ser capaz de captar relaciones sociales estables.

3. El principio del equilibrio en las interacciones

La teoría de la acción tiene como supuesto oculto el control que una persona ejerce sobre un objeto o una región del mundo: sin ese control, que debe ser necesariamente físico, la teoría de la acción no es más que una teoría de la representación de la decisión. Este control se traduce en una transformación del mundo: el resultado de la acción. Este resultado, más el supuesto del control nos permite reconstruir el proceso de decisión si tenemos suficientes datos acerca de las motivaciones o acerca de las creencias de la persona. Este es el proceso de interpretación al que nos remitimos en el caso de las conductas individuales (Broncano 1995). Pues bien, esta reconstrucción tiene un defecto básico cuando nos enfrentamos a una acción social: el resultado depende también de las expectativas sobre los otros individuos o, si se quiere, de las lecturas que cada uno hace de las mentes de los otros con el objeto de calcular cómo se realizará su aportación a la acción. Los deseos y creencias de los individuos son insuficientes para explicar la acción. El intérprete externo tiene que dilucidar la aportación de cada uno sobre la evidencia observable. ¿Cómo pasar desde la teoría de la acción individual al plano colectivo?. Desde Pareto sabemos que existe un correlato del Principio de Racionalidad: la noción de *equilibrio*.

Supongamos el caso más elemental de una interacción: un intercambio. Tomamos de Coleman, 1993 el análisis del intercambio, así como la teoría de la transferencia de autoridad. En esta situación comenzamos con dos personas, Sa y Sb que controlan respectivamente dos bienes A y B. Supongamos ahora que Sa desea B y que Sb desea A. Cada uno tendrá que emplear al otro como instrumento en la realización de su deseo. Hasta aquí tendríamos una teoría estándar de la acción individual, salvo que al representarse cada uno los medios para conseguir lo que buscan, deben representarse cada uno la mente del otro, dadas sus habilidades de sujetos sociales. Sin esta capacidad lo que ocurriría es que se trata al otro como un objeto más, lo que conlleva que, al menos en términos estadísticos, la imprevisibilidad de la conducta del otro haga imposible conseguir lo que queremos. Subrayamos de nuevo este carácter práctico de la *Teoría de la Mente*: reduce la imprevisibilidad de la conducta ajena y la convierte en potencialmente manipulable como instrumento de las motivaciones propias.

En ausencia de otras consideraciones o datos, la teoría de la racionalidad individual nos permite predecir que ambos agentes buscarán un intercambio de bienes. Pero no nos dice cómo se realizará este intercambio. La teoría del equilibrio sí nos permite hacerlo: nos indica que la cantidad de A que Sa transfiere a Sb llegará a un punto que tiene que ver con la cantidad de B que Sb transfiere a Sa. Este punto de equilibrio solamente está determinado por el valor mutuo de A en términos de B y viceversa, y por la cantidad de A y B que controlan cada uno de los agentes.

En la situación normal la teoría del equilibrio nos permite predecir de forma completa la interacción de intercambio. Suponemos que los actores son seres libres e independientes uno del otro: no mantienen ningún vínculo social. Y es precisamente este supuesto de independencia el que nos facilita la tarea interpretativa en contextos de interpretación radical para descubrir la existencia de relaciones sociales. Porque, a diferencia de lo que piensa el atomismo, no todas las acciones son acciones de intercambio bajo el supuesto de independencia. Por el contrario, las personas están atados unos a otros por múltiples relaciones de poder, de afecto, de competencia, etc. La desviación del equilibrio nos indica que está operando una relación por la cual uno de los agentes transfiere autoridad sobre A o B al otro sujeto. Esta transferencia se mide en términos de poder de Sa sobre Sb o viceversa. La explicación es múltiple: puede ser el afecto de Sb por Sa o, por el contrario, el miedo a Sa o pueden ser otras múltiples bases causales. Lo importante es que la desviación es estable (o estadísticamente significativa) y que es asimétrica.

Los humanos son seres sociales en un doble sentido: en primer lugar porque la representación de la mente del otro forma parte de su construcción interna de la decisión y en segundo lugar porque establecen vínculos de transferencia de autoridad permanente con respecto a los otros. La transferencia puede realizarse con personas concretas con quienes se

establecen vínculos afectivos (o de miedo, odio, envidia, etc.) o con otras personas en abstracto, como ocurre cuando se actúa por altruismo o por normas de equidad, y en general por normas de conducta que no distinguen entre individuos; también puede transferirse autoridad al grupo, como ocurre en los casos de acción colectiva para la producción de bienes públicos.

Es tarea de las ciencias sociales clasificar y explicar las formas que asumen estos patrones de transferencia de autoridad de cada agente a los otros: pueden ser normas, instituciones, sistemas de castigo o recompensa, lazos de familia o afecto, y cualesquiera otros de los que encontramos en las descripciones sociales. Pero ahora estamos en mejor situación para explicar la autonomía de lo social respecto a lo individual: las relaciones de transferencia de autoridad pueden estar realizadas de formas múltiples en los colectivos de agentes. Estudiando las desviaciones del equilibrio en los intercambios nos encontramos con una relación abstracta, que no explica los mecanismos causales, pero que nos facilita una información esencial sobre la interacción: nos informa precisamente de la existencia de una relación social.

El principio de equilibrio en las interacciones se convierte así en un principio de interpretación radical. Es sensible a las mínimas desviaciones y, cuando estas desviaciones se muestran sistemáticas, nos abre el espectro de relaciones sociales que están operando en el grupo. No hay nada misterioso en este principio, es un recurso habitual de todas las ciencias: el principio de inercia en mecánica clásica, por ejemplo, actúa metodológicamente de forma similar, a través de las desviaciones del momento nos permite inferir la existencia de fuerzas, que pueden tener sustentos causales diferentes, o la ley de Hardy-Weinberg en genética de poblaciones: las desviaciones de las distribuciones de caracteres esperadas nos permiten inferir la presencia de otras fuerzas como la selección natural, la deriva genética etc. O el Principio de Caridad en teoría de la acción individual, como ya hemos señalado insistentemente.

El carácter especial del principio de equilibrio no deriva de su *modus operandi* metodológico sino del hecho de que se apoya de forma esencial en la *Teoría de la Mente* como mecanismo psicológico constitutivo de los agentes. Si los agentes fueran autistas sus relaciones podrían explicarse conociendo sus apegos u otros mecanismos emotivos, pero no podríamos emplear el principio de equilibrio. Incluso si fuera cierto que nuestra constitución genética llevase incorporadas ciertas normas como un cierto sentido de la justicia distributiva, el principio de equilibrio seguiría siendo un recurso necesario para explicar los contextos en los que los sujetos aplican estas normas. Tampoco supone necesariamente una constitución egoísta o autointeresada. Imaginemos un sujeto que actuase de la manera máximamente altruista desde el punto de vista del sentido común, pero que debido a algún mecanismo interno supusiera que sus acciones son egoístas sistemáticamente: ¿podríamos calificarle como altruista?. Mi intuición es que no. Un altruista

debe saber que lo es, que actúa conscientemente desviándose del equilibrio previsible.

Es más, la Teoría de la Mente, el Principio de Equilibrio y el Principio de Racionalidad individual forman un conglomerado que cae o se sostiene junto: la capacidad de lectura de las mentes ajenas y su ocasional manipulación supone la racionalidad y una y otra llevan directamente al equilibrio en las interacciones con los otros. E inversamente, la existencia empírica de equilibrios en las interacciones, suponiendo la racionalidad, entraña la interpretación mutua de los agentes, o bien, suponiendo esta capacidad interpretativa, se concluye la racionalidad de los agentes. Y precisamente por ello las desviaciones del equilibrio son indicadores de relaciones estables o mecanismos de transferencia.

BIBLIOGRAFÍA

- Broncano, Fernando (1995) "La fuerza del PC (Principio de Caridad. Circunstancia, habilidad y contenido en la explicación de la acción" *Análisis filosófico* XV, 73-106(1996)
- Clark A. & D.Chalmers (1998) "The extended mind" *Analysis* 58, 7-19
- Clark A. (1998) "Where Brain, Body, and World Collide" *Daedalus* 127 257-280
- Carruthers, P & Smith, P (eds) (1996) *Theories of theories of mind*, Cambridge, Cambridge University Press
- Whiten, A. (ed) *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, Oxford, Basil Blackwell
- Byrne, R., A. Whiten, (eds) (1988) *Machiavellian Intelligence. Social Expertise and the evolution of Intellect in Monkeys, Apes and Humans*, Oxford, Oxford University Press
- Whiten, A., R. Byrne (eds) (1998) *Machiavellian Intelligence II. Extensions and Evaluations*, Cambridge, Cambridge University Press
- Coleman, James, (1993) *Foundations of Social Theory*, Cambridge MA, Harvard University Press
- Pettit, Philip (1993) *The Common Mind. An Essay on Psychology, Society and Politics*, Oxford, Oxford University Press

Fernando Broncano
Universidad de Salamanca
e-mail: broncano@gugu.usal.es

TEORÍA CAUSAL DE LA REFERENCIA Y REALISMO INTERNO

M^a CELESTE CANCELA SILVA
Universidad de Santiago de Compostela

ABSTRACT: Hilary Putnam presents his theory of reference in his famous work "The Meaning of «Meaning»" in 1975. As it is known, Putnam's thought changed from an externalist perspective to an internalist perspective. Such a change was a radical rupture with the ontological and epistemological assumptions of metaphysical realism. However, the semantic theory defended by Putnam in his second stage is very similar to the defended in the works of his first period.

My purpose in this contribution is to analyze the possible inconsistencies that can be derived from integrating Putnam's semantic theory with his internal realism, a situation that is associated to Putnam since 1976.

KEYWORDS: Metaphysical realism, internal realism, causal theory of reference, sense's theories, reference, sense.

1. Introducción

Comenzaré exponiendo, de una manera muy breve, las ideas clave del realismo interno a fin de poner de manifiesto los supuestos epistemológicos que se derivan de esta actitud filosófica y confrontarlos con la teoría de la referencia defendida por Hilary Putnam en los últimos años. Veremos cómo algunas tesis de la teoría causal de la referencia, pues este es el nombre que recibe la teoría semántica de Putnam, son inconsistentes con la epistemología característica o propia del realismo interno o, en otras palabras, veremos cómo fracasa el intento de querer combinar ambas teorías de Putnam, a saber, la teoría epistemológica y la semántica; ya que, como pretendo poner de manifiesto, de dicho intento surgen ciertas incoherencias incómodas para el realista interno. Terminaré este trabajo presentando cuáles de las tesis de la teoría semántica putnamiana serían consistentes con la epistemología que se deriva del realismo interno e intentaré reelaborar la teoría de la referencia de Putnam de tal modo que ya no resulte incompatible o inconsistente con este tipo de realismo.

2. Teoría semántica de Putnam

Podemos sintetizar, de manera muy esquemática, la doctrina del realismo interno en dos ideas fundamentales: por una parte, el supuesto que nos dice que no tiene sentido postular la existencia o inexistencia de un mundo independiente de la mente humana. Y, por otra parte, y como consecuencia de lo anterior, la tesis del relativismo conceptual que postula que los objetos que existen o las propiedades que predicamos de ellos son relativos a la elección de un esquema o marco conceptual.

Considero que las dos anteriores ideas, cruciales del realismo interno, son suficientes para atender a los objetivos principales de este trabajo. Teniendo estos supuestos internalistas presentes, voy a esbozar la teoría semántica de Putnam.

En los años setenta surge con Kripke, Putnam, Kaplan, Donnellan, Ruth Barcan Marcus entre otros, una teoría semántica conocida con el nombre de “teoría causal de la referencia”, “teoría de la referencia directa” o “nueva teoría de la referencia” emerge como una alternativa a las teorías del sentido de Frege, Strawson y Searle. La perspectiva que Putnam defendía en la época en que propone su teoría de la referencia era la del realismo metafísico; doctrina ésta que se ha convertido a partir de 1976 en el blanco privilegiado de sus críticas, las cuales, por otra parte, constituyeron los cimientos para su doctrina del realismo interno.

Aunque se produjo un cambio radical que bien podríamos calificar de metamorfosis en su pensamiento, sin embargo, su teoría semántica apenas sufrió los cambios que sería de esperar que acompañasen a su cambio de perspectiva.

Así, como la epistemología del realismo interno tiene su base en el rechazo rotundo de la perspectiva externalista, su teoría semántica tiene sus cimientos en el rechazo por parte de Putnam de las teorías del sentido. El objetivo principal de Putnam, en toda su obra, es mostrar cómo la referencia no viene determinada o fijada mediante nuestras representaciones mentales o nuestros estados psicológicos.

En la teoría semántica propuesta por Frege el sentido desempeña principalmente dos funciones: una función epistémica y una función de fijación de la referencia. Aquí me ocuparé de esta última función ya que es la que, según Putnam, plantea más problemas. Lo que quiere decir que el sentido desempeña una función de fijación de la extensión es que la referencia de una expresión viene determinada por el sentido que los sujetos asocian a dicha expresión. Putnam sostiene que esta tesis es falsa, si se entiende el sentido como aquello que el hablante tiene en su cabeza, esto es, como una representación mental y subjetiva; y argumenta esto recurriendo a varios experimentos mentales, en los que la referencia de términos como “olmo”, “haya”, “agua”, etc., no puede venir determinada por la representación mental

asociada a ellos. La propuesta de Putnam consiste en defender que la referencia se fija socialmente y que no depende, únicamente, de lo que los hablantes tienen en su cabeza; su propuesta se recoge perfectamente en su dictum: *los significados no están en la cabeza*.

Para mostrar cómo las extensiones de las expresiones se fijan socialmente Putnam recurre al fenómeno que él mismo ha denominado como *división del trabajo lingüístico*.

La tesis de la división del trabajo lingüístico se basa, nos dice Putnam, en la noción general de la división del trabajo. A partir de la división del trabajo surge la división de la tarea lingüística, los miembros de la sociedad desempeñan diferentes trabajos o funciones, y dichos miembros están familiarizados con un determinado vocabulario específico de su trabajo, llegando a ser unos expertos en el significado y la extensión de aquel vocabulario vinculado a su ámbito laboral. Así, no toda la comunidad lingüística tiene porqué saber cuál es el referente “exacto” de todas las expresiones, entre otras razones porque no está familiarizada con todo el vocabulario del cuerpo lingüístico. Son los diferentes especialistas los que mediante su trabajo nos dicen que, por ejemplo, el agua es H₂O o que un olmo es un árbol de la familia de las ulmáceas, que crece hasta la altura de 20 metros, que es de hoja caduca, etc. Es conveniente señalar que esto no significa que los “expertos” tengan un conocimiento más cabal de la referencia de todos los términos que el resto de los hablantes, sino que este personal especializado tiene mejor conocimiento de la referencia de algunos términos.

Yo considero, al igual que Putnam, que nuestra habilidad para usar el lenguaje viene determinada, en parte, por nuestros vínculos sociales con el resto de la comunidad lingüística; esto es, depende de nuestras interacciones con otros en una comunidad donde confluyen diversos intereses, capacidades y personas que podemos calificar de “expertos”. El planteamiento de Putnam acerca de términos como ‘fotones’, ‘cromosoma’, etc., es acertado: nosotros podemos hacer uso de este tipo de términos a pesar de nuestra ignorancia de esas cosas; ello es resultado de nuestros vínculos sociales con otros hablantes cuyo conocimiento de esos objetos es mucho más preciso que el nuestro. Lo que, en resumen, la tesis de la división del trabajo lingüístico nos viene a decir es que el lenguaje es un fenómeno social y no privado. Los hablantes de una lengua son igualmente capaces, en principio, de usar cada término del lenguaje en sus interacciones con el mundo, incluso cuando no todos los hablantes son capaces de relacionar todos los términos, que ellos usan, con el mundo.

Además de la tesis de la división de la tarea lingüística, Putnam también señala otro factor importante a la hora de determinar la referencia de un término y que las teorías del sentido no tuvieron en cuenta, dicho factor es la *contribución del entorno*. Este aspecto pone también de manifiesto que el

estado psicológico del hablante no fija la extensión de un término; ya que en lo que consista la extensión de un nombre va a depender de la sustancia a la que dicho nombre refiere. De esta manera no es correcto decir que el término “agua” usado por los hablantes de nuestro planeta y usado por los hablantes de la Tierra Gemela tiene la misma referencia, pues la sustancia a la que refiere dicho término en la Tierra es diferente de la sustancia a la que refiere el término “agua” en el planeta gemelo. Así, en 1750 tanto los hablantes terrícolas como sus hablantes gemelos estaban en el mismo estado psicológico respecto al término “agua” y, sin embargo, se estaban refiriendo a sustancias distintas (aunque, obviamente, ellos no lo supieran) cosa que no ocurriría si realmente el estado psicológico determinara la extensión.

3. Algunos problemas que se derivan de la teoría semántica de Putnam

Voy a analizar, en este apartado, algunas consecuencias que se derivan de cada una de las anteriores tesis semánticas de Putnam. Voy a empezar analizando la tesis de la contribución del entorno deteniéndome en el famoso ejemplo, ya citado en líneas anteriores, de las dos Tierras Gemelas que sólo difieren en el líquido que denominamos en nuestra Tierra mediante el término “agua”. Putnam lo que nos dice, de manera reiterada, mediante este ejemplo es lo siguiente: que una vez que hayamos determinado que el agua es H_2O , cuando nos refiramos a esta sustancia en cualquier mundo posible, donde dicha sustancia existe, nos estaremos refiriendo siempre a aquel líquido cuya naturaleza o composición es H_2O y no otra. Pero ya antes de que se produjese el desarrollo de la química que nos permitió conocer una nueva propiedad del agua, a saber, su composición, el término “agua” ya hacía referencia a la sustancia formada por la combinación de un volumen de oxígeno y dos de hidrógeno aunque nosotros no lo pudiéramos saber. En efecto, el agua siempre fue, es y siempre será H_2O aunque nosotros no lo hayamos sabido siempre. De esta manera, no es posible que el agua sea XYZ y no H_2O en alguna situación contrafáctica, aunque sí podría ser el caso que en algún mundo posible el agua no tuviera algunas de las propiedades superficiales como la de ser incolora o la de ser insípida.

Lo que se sigue de esto, desde mi punto de vista, es la suposición de un sustrato natural independiente de nuestro conocimiento que es el que, en definitiva, determina la identidad del referente a través de todos los mundos posibles. La propiedad esencial de una determinada sustancia es aquella que se preserva a través del tiempo y de todo marco conceptual y que nosotros, mediante al conocimiento científico, llegamos a conocer; esto es, la propiedad esencial pertenece a la sustancia independientemente de nuestro conocimiento de ello. Aquí aparece una de las principales dicotomías inadmisibles desde la perspectiva internalista. Desde el punto de vista del realismo interno, lo epistémico y lo ontológico están íntimamente relacionados. Lo oportuno, desde la filosofía del realismo interno, sería sostener que lo que sea la esencia

de algo no puede determinarse independientemente del esquema conceptual que escojamos; de tal modo que carecería de sentido hablar de propiedades esenciales al margen de todo marco conceptual.

Lo que estoy queriendo decir con todo esto es que no es posible predicar la existencia de propiedades esenciales al margen de todo esquema conceptual. Hay algo muy importante que Putnam no dice, y es que lo que sean las propiedades esenciales y las propiedades accidentales de un objeto dependen de los marcos conceptuales; esto es, no dependen exclusivamente del objeto. Así, dentro de un determinado esquema conceptual podemos considerar que, para que un líquido sea denominado “agua”, tenga unas determinadas propiedades, mientras que en un esquema conceptual distinto podemos fijarnos en otras propiedades diferentes para determinar si un cierto líquido puede considerarse o no agua.

En cuanto al fenómeno de la división de la tarea lingüística y en conexión con lo anterior, parece que Putnam reduce al conocimiento científico la absoluta competencia para determinar las propiedades esenciales de las sustancias. Así, por ejemplo, lo esencial del agua para *ser* agua es estar constituida por H₂O, y lo esencial de los limones para *ser* limones es poseer una determinada estructura cromosómica. Se produce, de este modo, una especialización en la que un determinado esquema conceptual tiene preferencia a la hora de determinar cuál es la propiedad esencial de una sustancia. Una de las ideas fundamentales de la tesis del relativismo conceptual es que la pregunta ¿de qué objetos consta el mundo? depende del esquema conceptual dentro del cual formulemos la pregunta, de tal forma que la respuesta que demos a dicha pregunta va a venir, en parte, determinada por el marco conceptual. Sin embargo, si nos atenemos a la tesis de la división del trabajo lingüístico, tal y como la expone Putnam, parece que la respuesta a la anterior pregunta va a depender, casi exclusivamente, de lo que nos diga la ciencia.

4. Conclusiones

En resumen, considero que las tesis de la contribución del entorno y de la división del trabajo lingüístico son claves para cualquier planteamiento posible de una teoría semántica; pero si queremos ser fieles a la perspectiva del realismo interno, hay que despojar a estas tesis tanto del esencialismo como de la especialización que en ellas subyace. De esta forma, pienso que dentro de un determinado esquema conceptual podemos considerar que una cierta sustancia u objeto para *ser* dicha sustancia u objeto posea unas determinadas propiedades, mientras que en un esquema conceptual distinto podemos fijarnos en otras propiedades diferentes. También considero importante eliminar la idea de que hay un esquema conceptual “privilegiado” o mejor que los demás o sobre el cual pivote la determinación de lo que es la propiedad esencial de algo. Con esto no estoy queriendo decir que todos los

esquemas conceptuales son igualmente buenos o que “todo vale”, sino que no podemos reducir todos los puntos de vista a una sola perspectiva, como parece que se deriva de los textos semánticos de Putnam.

BIBLIOGRAFÍA

- Hierro S. Pescador, J., *Principios de filosofía del lenguaje*, Alianza Editorial, Madrid.
Putnam, H., *Razón, verdad e historia*. Ed. Tecnos, Madrid, 1988.
Putnam, H., *Representación y realidad*, Ed. Gedisa, Barcelona, 1995.
Schwartz, Stephen P. (Ed.), *Naming, Necessity and Natural Kinds*, Cornell University Press, New York, 1977.
Schartz, Stephen P., “Natural Kinds and Nominal Kinds”, en *Mind* 89: 182-195, 1980.
Valdés Villanueva, Luis Ml., *La búsqueda del Significado*, Ed. Tecnos, Madrid, 1991.

M^a Celeste Cancela Silva
Univ. de Santiago de Compostela
e-mail: mceleste@usc.es



PROBLEMAS DE LA DEFINICIÓN DE FUNCIÓN BIOLÓGICA

ARANTZA ETXEBERRIA

Euskal Herriko Unibertsitatea - Universidad del País Vasco

ABSTRACT: Three main accounts of biological function in the literature are discussed in this paper: the etiological function, the dispositional function and the causal-role (Cummins) function. They are all said to be “analytic”, in that they proceed through some division of a whole in parts, so that the function of each item is defined, either in historical/evolutionary or dispositional terms. In contrast with this perspective, a “synthetic” one is outlined, with the aim to sketch a concept of the origins of biological functions and their role in the autonomous living organization.

KEYWORDS: Autonomy, causal-role function, dispositional function, etiological function.

Introducción

El análisis filosófico del concepto de “función” crea verdaderos quebraderos de cabeza. Por un lado, por su problemática relación con las causas: aunque una función pueda definirse como el rol o el papel causal de un determinado elemento o particular (la función de algo es causar un efecto), la función de un elemento no puede determinarse atendiendo exclusivamente a su estructura; por otro, por su relación con la intención o la teleología, de las que a menudo la función se hace derivar, aunque en el orden natural parecería más plausible pensar que ésta deba preceder a cualquier intención, consciente o no. La cuestión a plantear es si hay alguna perspectiva desde la que se pueda dar cuenta de las funciones como el producto de ciertas relaciones causales entre elementos del sistema, es decir, un tipo especial de propiedad relacional, sin caer en posiciones eliminativistas.

En estos últimos años una vertiente de los partidarios/as del funcionalismo dice que un concepto de función basado en la función biológica es más adecuado que el que deriva de la máquina de Turing para tratar problemas de otros ámbitos en los que se apela a este concepto, como, por

ejemplo, la psicología cognitiva (Sober, 1985). Aún así, no está claro cuál de los posibles conceptos de función biológica deba elegirse para tal tarea de reformulación del funcionalismo. La mayor parte de sus proponentes han optado por una versión histórica (ver, por ejemplo: Millikan, 1989; Godfrey-Smith, 1996), pero debe tenerse en cuenta que en biología todavía no hay un consenso sobre esta cuestión (aunque a veces parezca que quienes reducen la biología a la biología evolutiva han decidido ya).

En general, el concepto de función biológica está estrechamente relacionado con el de adaptación de los organismos a sus entornos. Este concepto tiene dos sentidos en la biología evolutiva reciente, un sentido ahistórico y otro histórico (Burian, 1983). En el primer sentido, la adaptación implica que un cierto rasgo supone algún tipo de ventaja o contribuye a la solución de algún problema. Por ejemplo: la ligereza hace que el ciervo esté mejor adaptado si, en igualdad de condiciones, contribuye a la solución de un problema planteado al ciervo, por ejemplo, escapar de un depredador. En el segundo sentido, histórico, la adaptación significa que un rasgo es producto de un proceso en el que éste ha sido seleccionado. En este caso la ligereza es una adaptación del ciervo si ha sido moldeada por un proceso histórico en el que una relativa ligereza de ciervos anteriores permitió dar forma a la ligereza del ciervo actual. Para algunos biólogos como Gould (1983), la generalización de este segundo sentido en la biología evolutiva es una consecuencia del "endurecimiento" de la teoría sintética a partir de los años sesenta, que reduce todas las fuerzas evolutivas a una, la selección natural, y sus críticas le han llevado a proponer otros conceptos, como el de exaptación, para referirse a los casos en los que un rasgo sea adaptativo en el primer sentido, pero no sea el producto de una evolución por selección natural (Gould & Vrba, 1982).

En concordancia con la manera histórica de entender la adaptación (ver, por ejemplo, Williams, 1966), se formula en los años 70 una versión afin del concepto de función biológica. Su desarrollo filosófico se produce por dos motivaciones principales: por un lado, la aparente posibilidad de eliminar la teleología en la aclaración de este concepto (frente a otros tratamientos que conciben la función como una propiedad del comportamiento dirigido a objetivos) y, por otro, la naturalización del concepto mediante la apelación a una teoría científica que dé cuenta de él: la teoría de la evolución (se ha discutido si ello supone o no reemplazar al análisis conceptual (ver Neander, 1991)). Sin embargo, aunque tales consecuencias sean sin duda deseables, parecen ser inherentes a cualquier concepto biológico de función, por lo que no está del todo claro que sólo un concepto histórico o evolutivo pueda generarlas.

En este trabajo me ocupo de discutir el tratamiento más adecuado de la función biológica (aunque no llegaré a extraer conclusiones sobre sus consecuencias para las premisas generales del funcionalismo). Tras repasar críticamente las versiones más habituales (que llamo “analíticas”), propongo la consideración de una noción “sintética” de función, implícita en desarrollos de teorías de la autonomía biológica, que podría llevar a proponer una definición alternativa en trabajos sucesivos. Trato, así, de caracterizar un cambio de perspectiva: de las funciones de los elementos particulares a las propiedades relacionales inmersas en entidades autónomas, y apuntar posibles consecuencias del mismo.

Función: concepciones analíticas

Las acepciones en las que el concepto de función se ha usado son muchas, Wimsatt (1972) distingue hasta diez diferentes. Aún así, la discusión filosófica más reciente (ver Allen, Beckoff & Lauder, 1998 para una antología) se ha centrado mayormente en tres alternativas: la función etiológica, la función disposicional y la función de Cummins. El primer y el tercer enfoque se consideran seminales en la literatura (son también llamados, respectivamente, la explicación de los “efectos seleccionados” y del “rol causal” en Neander (1991)). El primero presenta una noción causal de la función y el tercero, uno derivado de las características de los sistemas complejos. Mientras el primero ha sido posteriormente reinterpretado en los términos históricos de la biología evolutiva, el tercero se adecúa más al tipo de análisis propio de la morfología o la fisiología. La función disposicional discute el planteamiento exclusivamente histórico de la función etiológica, que critica explícitamente, y caracteriza las funciones como propiedades disposicionales que dan cuenta de los rasgos adaptativos (en el sentido ahistórico que mencionábamos en la introducción).

Considero que estas definiciones son “analíticas”, porque tratan el problema de la función de los elementos sin considerar las relaciones entre los mismos y presuponiendo, en el caso de las funciones de los organismos, un supuesto análisis funcional previo del mismo, producto del cuál son los diferentes rasgos y sus funciones adaptativas, que hay que dilucidar. Esta estrategia es propia del análisis funcional, pero tampoco es ajena a la biología evolutiva donde la adaptación se estudia dividiendo al organismo en múltiples partes, cuya utilidad será evidente: las patas del topo están adaptadas a la excavación de galerías, las gibas del camello constituyen una reserva de agua que le permite vivir en el desierto, la trompa del elefante le sirve para coger los alimentos, la constitución de sus patas permite sostener su cuerpo, etc. En definitiva, los caracteres se estudian por separado y desde el punto de vista de

su utilidad para el organismo. Una consecuencia de este enfoque es que el esfuerzo teórico está dirigido a especificar la función de los elementos en relación con la adaptación, en detrimento de una posible concepción alternativa que tratara de encontrar qué tipo de relaciones en el sistema hacen que un cierto elemento sea funcional en el mismo.

La función etiológica

Wright (1973) propone una función etiológica o causal de la función para alejarse de una noción teleológica (goal-directed):

The function of X is Z means that 1) X is there because it does Z and 2) Z is a consequence of X being there. (Wright (1973), reprinted in Sober, 1984, p. 363)

Esta noción es etiológica o causal porque explica la función de un elemento en términos de la razón de la existencia o presencia de ese elemento. La idea es que hay una diferencia entre decir para qué es un elemento o por qué está ahí y, por consiguiente, una adscripción funcional basada en el enunciado 1) es causal, porque lo que explica es la presencia de X, mientras que las consecuencias de la estructura de X son un aspecto secundario para definir función. Así, el elemento no es funcional por lo que hace, sino por la intención consciente o el suceso evolutivo (selección natural) que es la razón de que esté ahí.

Wright se proponía diferentes cosas mediante esta versión. Por un lado, que la definición se pueda aplicar a cosas que no tengan comportamiento (una silla, un pivote que mantiene abierta una puerta), pero que son habitualmente consideradas como funcionales. Además, que se pueda especificar unívocamente la función de un elemento, es decir, distinguir *la* función de algo de otras funciones posibles que ese algo pueda tener, pero no son la razón de su existencia (mi baffle sostiene el teléfono, pero no esa es *su* función). También distinguir la función de algo de otros efectos accidentales posibles del elemento (el ritmo del corazón de una persona alerta al médico de una enfermedad, pero esa no es la razón por la que la persona tiene corazón). Por último, otro de sus cometidos era que el análisis del concepto pudiera servir igualmente para las funciones conscientes y las naturales, equiparando en bastante grado el valor explicativo de la intención consciente y de la selección natural.

Las versiones más recientes de esta explicación de las funciones (a veces llamada la teoría de los “efectos seleccionados”) limitan el análisis al ámbito de la biología evolutiva y a las funciones adaptativas generadas por selección natural (Millikan, 1989; Neander, 1991). La noción de “función

propia" de Millikan es una variante de ésta con un compromiso muy fuerte por el hecho de que se trata de una definición teórica de la noción.

La función disposicional

Este enfoque es ahistórico, identifica la función de un elemento o un rasgo con alguna de las propiedades causales en vigor, en general, aquella que tiene algún significado evolutivo para la reproducción o la supervivencia del organismo.

La discusión filosófica más enérgica de esta dirección es la de Bigelow & Pargetter (1987), que critican la función etiológica por estar basada en un retroceso explicativo (backward-looking). En su lugar proponen una función dirigida hacia adelante (forward-looking), a partir de un análisis disposicional de las propiedades. Una propiedad disposicional de una entidad funcional (un rasgo o un carácter) confiere cierta propensión a sobrevivir bajo ciertas condiciones. Esta propensión puede analizarse como una propiedad subjuntiva que especifica qué sucederá o qué es probable que suceda en circunstancias determinadas del futuro.

Las propensiones funcionales se analizan de la misma manera que la eficacia (fitness) se analizaba en Mills & Beatty (1984). Ahí se trataba de salvar a la teoría evolutiva de la acusación de tautología, especificando los criterios objetivos por los que un rasgo dado es eficaz (en lugar de dejar que la eficacia se defina, circularmente, por la tasa de supervivencia). Por otro lado, frente a las críticas eliminativistas que argumentan que las funciones se refieren a efectos futuros que podrían no darse, se dice que la disposición de un rasgo es causal:

"So a character or structure has a certain function when it has the propensity for selection in virtue of that character or structure's having the relevant effects" (Bigelow & Pargetter, 1987, p. 194).

Para los disposicionalistas el análisis etiológico dice que un carácter tiene una función si contribuye a la supervivencia en el pasado, mientras que ellos pueden considerar los efectos futuros:

"The aetiological theory describes a character *now* as serving a function when it did confer propensities that improved the chances of survival. We suggest that it is appropriate, in such a case, to say that the character has been serving that function all along". (Bigelow & Pargetter, 1987, p. 192)

La versión disposicional es interesante porque toma en cuenta la naturaleza del carácter y no sólo el hecho de que haya sido seleccionado, pero

no trata de relacionar tal propensión con algún tipo de estructura. La función es, por ello, ontológicamente superveniente con respecto a la estructura.

La función de Cummins (o del rol causal)

Cummins (1976) propuso una definición de función desde una perspectiva completamente diferente, de hecho hoy en día pocos/as investigadores consideran que sea incompatible con alguna de las anteriores y muchos aceptan los dos sentidos. Es una noción especial porque no apela a propósito alguno, sea actual o evolutivo (Amundson & Lauder, 1993-4).

Basa su explicación en el análisis funcional, mediante el que se explica una capacidad del sistema apelando a las capacidades de sus partes componentes. Estas capacidades no son propósitos o fines del sistema, son elegidas por el/la observador, quien trata de armar explicaciones de cómo pueden surgir tales capacidades a partir de las partes componentes. Las funciones que se asignan a las partes son, pues, relativas tanto a la capacidad global escogida para el análisis, como a la explicación funcional construida:

x functions as a F in s (or: the function of x in s is to F) relative to an analytical account A of s's capacity to G just in case x is capable of F-ing in s and A appropriately and adequately accounts for s's capacity to G by, in part, appealing to the capacity of x to F in s. (Cummins (1975) en Sober (1983), pag. 405).

Algunos problemas de estas concepciones

Las definiciones esbozadas hasta aquí generan muchas discusión entre las personas partidarias de cada una y por parte de defensores de otras perspectivas. Algunos de los problemas son el papel de la historia, la relevancia de la estructura material, la naturaleza de la selección natural y la importancia de la organización.

¿Cuál es el papel de la historia?

La primera cuestión es la disputa sobre las explicaciones históricas: mientras para algunos sólo la historia (reciente: ver Godfrey-Smith (1994) y Kitcher (1993)) justifica las funciones, otros consideran que esa vuelta hacia atrás trae consigo consecuencias indeseadas, como el mencionado problema de las exaptaciones, o referentes a la historia del uso del concepto (antes de que se formulara la teoría de la evolución por selección natural ¿había funciones?) o, incluso, metafísicamente perversas:

“Consider the possible world identical to this one in all matters of laws and particular matters of fact, except that it came into existence by chance (or without cause) five minutes ago. Now, on even the over-arching theory there are now functions; ...” Bigelow and Pargetter (1987), pag. 188

De forma más general también hay debate sobre el papel de la evolución a la hora de explicar la fenomenología biológica: mientras evolucionistas como Dobzhansky afirman que “nada en biología tiene sentido si no es a la luz de la evolución”, biólogos teóricos como Rosen (1991) o Maturana y Varela (1984) pueden prescindir de ella:

“To me it is easy to conceive of life, and hence of biology, without evolution. But not of evolution without life. Thus evolution is a corollary of the living, the consequence of specialized somatic activities, and not the other way around. Indeed, it may very well be more a property of particular realizations of life, rather than of life itself.” (Rosen (1991), pag. 255

pues consideran que la evolución poco o nada puede decir sobre la cuestión más acuciante que plantea la biología, que es la explicación de la organización biológica.

¿Tiene importancia la estructura?

Otra crítica a la función etiológica pide que la noción de función explicité qué es lo que el elemento funcional hace, la manera específica como el elemento produce su efecto ahora (Brandon, 1990; Amundson & Lauder, 1993). Entre otras razones porque es el sentido en el que la mayor parte de la biología y la medicina se refieren a las entidades funcionales.

Además, una versión naturalista de las funciones tiene que poder dar alguna pista sobre su origen en el mundo natural, algo que la selección natural no puede, ésta sólo describe propiedades medias, no puede explicar la existencia de rasgos concretos en los individuos (Sober, 1995). Este problema ha hecho que Salthe se refiera a las explicaciones darwinistas como semióticas (Salthe, 1998), mientras que cada vez es más patente la relevancia de las aportaciones de las teorías autoorganizativas a la teoría de la evolución (ver: Maynard Smith et al, 1985; Depew & Weber, 1995; Kauffman, 1993).

Por ello, nos podemos preguntar si la explicación de la función de un elemento, aislado de su contexto en el organismo, es suficiente. Este problema afecta tanto a la función etiológica como a la disposicional, no así a la de Cummins. De todas maneras, incluso ésta, por ser analítica, debería

complementarse con un concepto sintético de la aparición de funciones en el conjunto dinámico de componentes que forman una organización viva.

¿Es la selección natural una clase especial de diseño?

La selección natural es a veces equiparada con la intención consciente, incluso aunque se sepa que la evolución es ciega y la selección no actúa más que sobre una variabilidad aleatoria. Tal equiparación, explícita, como hemos dicho, en Wright, siempre produce confusión, incluso entre los evolucionistas:

“(S)uch attributions of goals to evolutionary consequences represents a category mistake -a confusion of two modes of discourse which, though they exhibit strong analogies, are fundamentally different. I think that the fallacy involves a kind of “reverse Darwinian” move. Darwin inferred the concept of “natural selection” by analogy with the “artificial selection” of breeders. I suspect that the term “natural selection” still has some of these “selective” connotations. Interestingly the term preferred by biologists such as Lewontin (1969), namely “differential reproduction”, lacks those connotations...” (Schaffner, 1993, p. 374)

La limitación de las fuerzas evolutivas a la selección natural hace que el peligro de asimilar la selección con una fuerza omnipotente que puede “crear” cualquier forma adaptada esté más cercano.

¿Qué pasa con la organización?

Una explicación biológica aceptable del concepto de organización requiere explicar las condiciones dinámicas y materiales de capacidades que dan cuenta de su autonomía, como la autoorganización y el automantenimiento. Para ello no basta con la explicación aislada de las funciones de los componentes que las forman, los modelos autoorganizativos suelen mostrar que las capacidades que nos interesan están distribuidas en redes de componentes y no pueden ser asignadas trivialmente a partes de los mismos. Por ello, tratar de desarrollar un concepto sintético de función biológica que pueda aplicarse para los modelos de sistemas autoorganizados y autónomos es un desafío que la filosofía debería tomar para sí.

Función y autonomía: concepciones sintéticas

Muchos de los trabajos que estudian la función biológica desde el punto de vista sintético de su aparición son eliminativistas (por ejemplo, Varela (1979) considera que las funciones son descripciones del observador/a

que nada tienen que ver con la descripción operacional de la dinámica propia del sistema) o, en general, escépticos sobre la posibilidad de encontrar procedimientos que permitan identificar unívocamente las funciones. Eso no significa que la investigación filosófica deba conformarse con la discusión de los resultados obtenidos hasta el momento, es preciso idear nuevas formas de estudiar el problema.

El problema de la definición de función biológica puede compararse con el problema de la representación en ciencia cognitiva: las teorías existentes son insatisfactorias, pero renunciar al concepto resulta demasiado costoso y tal vez sea insensato. Lo que se requiere es una reconceptualización que integre los desarrollos teóricos y los descubrimientos empíricos de la ciencia.

Este estudio puede abordarse sintéticamente, desde el punto de vista de su aparición o emergencia. Ni la versión etiológica ni la disposicional pueden explicar el origen de las funciones en la naturaleza. La primera deja este problema a la selección natural, que tiene problemas para explicar episodios de origen, la segunda trata de dotar a las funciones de un contenido disposicional, pero las funciones son propiedades que difícilmente se pueden especificar de antemano. Cuando la biología teórica se enfrenta a problemas del origen o la creación de funciones, las definiciones anteriormente consideradas son de poca utilidad.

Por tanto, si las funciones no tienen contenido constitutivo (sea definido histórica o disposicionalmente), tal vez sea mejor definir las funciones como propiedades relacionales que establecen algún tipo muy especial de relación dentro del sistema, una relación que tienen que ver con la formación de un entidad autónoma. Pattee (1982) ha tratado de analizar este tipo de relaciones como el establecimiento autónomo de constricciones de medición y de control en el sistema que establecen un cierre funcional autónomo. Esto tiene una serie de consecuencias paradójicas para el problema de la función biológica: por un lado la función surge del entramado de relaciones que hacen posible el cierre, porque los componentes adquieren sus papeles en la dinámica que genera tal cierre; por otro, el hecho de que el sistema sea cerrado hace que la identificación de algo como funcional sea sólo provisional y supeditada a las condiciones propias de la autonomía.

¿Sería esta una conclusión escéptica o eliminativista respecto de la necesidad y posibilidad de definir funciones? No necesariamente. El estudio de este tipo de relaciones es imprescindible para el análisis filosófico de los modelos de organización biológica que se construyen para estudiar el origen de la vida, el metabolismo o el sistema inmunológico.

Conclusión

El concepto de función puede tratarse analíticamente (como en general lo ha hecho la filosofía de corte analítico) o sintéticamente (como es característico de la biología teórica y la vida artificial, pero está insuficientemente desarrollado conceptualmente). El tratamiento sintético es, quizás, complementario del analítico, pero desvela muchos problemas y asunciones que encierra éste último. La diferencia fundamental de ambos tratamientos es que el primero va a tratar de conferir al concepto de función de un contenido constitutivo (la función de F es ...), mientras que el segundo trata de entender en que consiste la participación funcional de algo en un sistema, en contraste con su participación causal, y cómo puede surgir tal participación. Podríamos decir que mientras el primero es un tratamiento "semántico" de la función, el segundo es un tratamiento "ecológico".

Es importante tener en cuenta que el concepto de función caracteriza el trabajo de la biología, y lo hace distinto del de la física (causa) y el de las ciencias de lo mental (representación), de forma que la relación entre las tres ciencias es tan discontinua como las discontinuidades generadas por sus conceptos centrales. La tradición que considera que las funciones son "supervenientes" hace la elección ontológica de considerar que las entidades "reales" se sitúan a un nivel "fundamental" que coincide con el de las partículas físicas elementales. Sin embargo, poniendo en cuestión la realidad de las funciones biológicas, también se duda sobre las entidades constituidas por ellas (y ¿hay algo que nos importe más?).

Reconocimiento

Este trabajo ha sido posible gracias a la ayuda recibida por los proyectos DGICYT PB95-0502 del Ministerio de Educación y Ciencia y HU-1998-142 y EX-1998-146 del Gobierno Vasco-Eusko Jaurilaritza.

BIBLIOGRAFÍA

- Allen, C., Beckoff, M. & Lauder, G. (1998) *Nature's purposes. Analyses of function and design in Biology*, Cambridge, MA & London: MIT Press.
- Amundson, R. & Lauder, G. V. (1993-4) "Function without purpose: The uses of causal role function in evolutionary biology", *Biology & Philosophy*, 9: 443-69. (Reimpreso en D. Hull & M. Ruse (eds) *The Philosophy of Biology*, Oxford University Press, 1998, pags. 227-257.)
- Bigelow, J. & Pargetter, R. (1987) "Functions", *The Journal of Philosophy* 84: 181-196.
- Brandon, R.N. (1990) *Adaptation and Environment*, Princeton University Press.
- Burian, R. (1983) "Adaptation". En M. Grene Ed. *Dimensions of Darwinism*, Cambridge University Press, pags. 287-314.
- Cummins, R. (1975) "Functional Analysis", *The Journal of Philosophy* 72: 741-765. (Reimpreso en Sober, E. (ed.) *Conceptual Issues in Evolutionary Biology*, MIT Press, 1984, pags. 386-407).

- Depew, D. & Weber, B. (1995) *Darwinism Evolving. Systems Dynamics and the Genealogy of Natural Selection*, Cambridge, MA & London: MIT Press.
- Gould, S.J. (1983) "The hardening of the modern synthesis". En M. Grene Ed. *Dimensions of Darwinism*, Cambridge University Press, pages. 71-93.
- Gould, S.J. & Lewontin, R. (1979) "The spandrels of San Marco and the Panglossian Paradigm: A critique of the Adaptationist Program", *Proceedings of the Royal Society of London*, CCV: 581-98.
- Gould, S.J. & Vrba, E.S. (1982) "Exaptation—A missing term in the Science of Form", *Paleobiology* 8/1: 4-15. (Reimpreso en D. Hull & M. Ruse (eds) *The Philosophy of Biology*, Oxford University Press, 1998, pages. 52-71).
- Kauffman, S.A. (1993) *The Origins of Order: Self-organization and Selection in Evolution*, Oxford University Press.
- Kitcher, P. (1993) "Function and Design", *Midwest Studies in Philosophy*, 18: 379-97. (Reimpreso en D. Hull & M. Ruse (eds) *The Philosophy of Biology*, Oxford University Press, 1998: 258-279.
- Godfrey-Smith, P. (1993) "Functions: Consensus without unity", *Pacific Philosophical Quarterly*, 74: 196-208. (Reimpreso en D. Hull & M. Ruse (eds) *The Philosophy of Biology*, Oxford University Press, 1998, pages. 280-292.)
- Godfrey-Smith, P. (1994) "A Modern History Theory of Functions", *Noûs*, 28: 344-362.
- Godfrey-Smith, P. (1996) *Complexity and the Function of Mind in Nature*, Cambridge University Press.
- Maturana, H. & Varela, F. (1984) *El árbol del conocimiento*, Santiago de Chile: Editorial Universitaria.
- Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Laude, R., Raup, D. & Wolpert, L. (1985) "Developmental constraints and evolution: A perspective from the Mountain Lake Conference on development and evolution", *Quarterly reviews of Biology*, 60: 265-287.
- Millikan, R. (1989) "In defense of proper Functions", *Philosophy of Science* 56:288-302. (Reimpreso en Millikan, R. *White Queen Psychology and other essays for Alice*, MIT Press, 1993.).
- Mills, S. & Beatty, J. (1984) "The Propensity Interpretation of Fitness". (Reimpreso en *Conceptual Issues in Evolutionary Biology*, E. Sober (ed.), MIT Press, pages. 37-57.)
- Neander, K. (1991) "Functions as selected effects: The conceptual analyst's defense", *Philosophy of Science*, 58: 168-184.
- Pattee H. H. (1982) "Cell Psychology: An Evolutionary Approach to the Symbol-Matter Problem", *Cognition and Brain Theory* 5 (4): 325-341.
- Rosen, R. (1991) *Life itself*, Columbia University Press.
- Rosenberg, A. (1985) *The structure of Biological Science*, Cambridge University Press.
- Salthe, S.N. (1998) "The role of natural selection theory in understanding evolutionary systems". En G. Van der Vijver, G., S. Salthe, S. & M. Delpo Eds. *Evolutionary Systems. Biological and Epistemological Perspectives on Selection and Self-organization*, Dordrecht: Kluwer, pages. 13-20.
- Schaffner, K. (1993) *Discovery and Explanation in Biology and Medicine*, University of Chicago Press.
- Sober, E. (1985) "Panglossian Functionalism and the Philosophy of Mind", *Synthese*, 64: 165-193.
- Sober, E. (1995) "Natural selection and distributive explanation: A reply to Neander", *British Journal for the Philosophy of Science* 46: 384-397.
- Varela, F. (1979) *Principles of Biological Autonomy*, Elsevier North Holland, New York.
- Williams, G.C. (1966) *Adaptation and Natural Selection*, Princeton University Press.
- Wimsatt, W. (1972) "Teleology and the Logical Structure of Function Statements", *Studies in the History and Philosophy of Science* 3: 1-80.
- Wright, L. (1973) "Functions", *The Philosophical Review*, 82: 139-168. (Reimpreso en Sober, E. (ed.) *Conceptual Issues in Evolutionary Biology*, MIT Press, 1984:347-368).

Arantza Etxeberria

Euskal Herriko Unibertsitatea

e-mail: ylbetaga@sf.ehu.es



SEIS O SIETE ARGUMENTOS EN PRO DEL PRINCIPIO DE CAUSALIDAD

MARÍA JOSÉ GARCÍA ENCINAS
Rutgers University

ABSTRACT: The paper considers some historical and contemporary a priori arguments in favour of the principle of causation and argues that none of them reaches its goal.

KEYWORDS: A priori, causality, causality principle.

1. Introducción: el principio de causalidad

"Ex nihilo, nihil fit" o "Si algo ocurre entonces tiene una causa" son probablemente las más acertadas formulaciones del principio de causalidad. También usuales son: "Necesariamente, todo tiene una causa" o "Todo, necesariamente, tiene una causa". La relación entre éstas y las anteriores no es difícil de ver:

- Si es cierto que "Necesariamente, todo (lo que ocurre) tiene una causa" entonces el principio de causalidad es cierto.
- Si es cierto que "Todo (lo que ocurre), necesariamente, tiene una causa" entonces es cierto que "Necesariamente, todo (lo que ocurre) tiene una causa".

Aún es posible considerar otras formulaciones del principio: "Nada en la tierra ocurre sin una causa" o "Nada puede existir ni dejar de existir sin una causa" o "Todo lo que comienza a existir nace necesariamente de alguna causa" o "Necesariamente, todo lo que tiene un principio tiene una causa"¹. Si la verdad de estas afirmaciones conlleva o no la verdad del principio de causalidad, depende claramente de qué sea una ocurrencia, de si lo que ocurre debe suceder "en la tierra" o "en el tiempo" o "tener un comienzo de

¹ La mayor parte de estas fórmulas pueden leerse en Bunge (1961, p. 4), aunque en su terminología se presentan como diferentes formulaciones de su "causalismo".

existencia". La relevancia de estas restricciones a la formulación del principio se hará evidente enseguida. En concreto, se verá cómo algunos filósofos las han preferido cuando, queriendo defender el principio de causalidad, pretendían al tiempo admitir como cierto lo que, en principio, podría ser excepcional al mismo, como la existencia de una primera causa o de Dios.

Comúnmente, la certeza del principio, o se ha justificado por razones prácticas (como por su utilidad para el razonamiento ordinario o incluso científico), o se ha defendido apelando a nuestras intuiciones más básicas. Por ejemplo, Th. Reid escribía:

That neither existence, nor any mode of existence, can begin without a cause is a principle that appears very early in the mind of a man; and it is so universal, and so firmly rooted in human nature, that the most determined scepticism cannot eradicate it².

Aunque este tipo de razones son habituales para la manutención del principio, no han faltado quienes han pretendido la *demonstración* de su certeza. Son estos los que se consideran en las páginas que siguen. La intención: algo más que mostrar que ninguno de ellos cumple su objetivo.

2. Locke: no es posible que la nada cause

Son varios los filósofos que han argumentado la existencia de Dios, como primera causa, desde la aceptación del principio de causalidad. Dentro de estos, además, están los que han querido primero defender el principio. Un ejemplo podría ser Locke:

No hay verdad más evidente que la de que tiene que haber algo desde la eternidad. Jamás he oído a nadie tan sin razón, ni a nadie que pudiera suponer una contradicción tan manifiesta, como para admitir un tiempo en el que hubiera una nada perfecta. Porque sería el mayor de todos los absurdos imaginar que la pura nada, la negación perfecta y ausencia de todo ser, pudiera alguna vez producir alguna existencia real³.

Locke parece defender que [(2)] desde siempre, todo tiene una causa, desde la premisa de que [(1)] la nada (o ausencia de ser) no puede ser causa del ser. Es decir, *puesto que es absurdo admitir que la nada pueda causar algo entonces, desde siempre, todo proviene de algo*.

¿Es aceptable este razonamiento como demostrativo del principio de causalidad? Aceptado (1) como verdadero, ¿podemos concluir (2)? Parece que no a menos que de alguna manera se asuma a su vez una premisa de la forma

² Reid (1969), p. 267.

³ Locke (1690), b. iv, ch. 10 (p. 624)

(P): Si “algo” no es causado por “algo”, entonces “algo” es causado por “nada”. O no, a menos que (1) se lea al modo del viejo dicho “de donde no hay no se puede sacar”; i.e. se asuma que todo lo que hay *tiene que provenir* (de lo que ya hay). De cualquier manera se seguiría que necesariamente “algo” es causado por “algo” y, por tanto que (2) “desde siempre, “algo” es causado que “algo””. Pero tanto (P) como esta lectura de (1) suponen que todo lo que ocurre es causado. Para aceptar el razonamiento de Locke, por tanto, se ha de asumir el principio de causalidad⁴.

(P) Si “algo” no es causado por “algo”, entonces “algo” es causado por “nada”

(1) No es posible que “nada” cause “algo” <---> “Algo” tiene que ser causado por “algo”⁵

(2) Desde siempre, “algo” causa “algo”

3. Descartes: la realidad del efecto está contenida en la causa

También Descartes, de camino en su argumento para demostrar la existencia de un Dios garante de todo conocimiento sensible, parece defender el apriorístico conocimiento del principio de causalidad:

[E]s cosa manifiesta, en virtud de la luz natural, que debe haber por lo menos tanta realidad en la causa eficiente y total como en su efecto: pues ¿de dónde puede sacar el efecto su realidad, si no es de la causa? ¿Y cómo podría esa causa comunicársela, si no la tuviera ella misma?

Y de ahí se sigue, no sólo que la nada no podría producir cosa alguna, sino que lo más perfecto, es decir, lo que contiene más realidad, no puede provenir de lo menos perfecto (...) Por ejemplo, la piedra que aún no existe no puede empezar a existir ahora si no es producida por algo que tenga en sí formalmente o eminentemente todo lo que entra en la composición de la piedra⁶.

Descartes presenta el principio de causalidad bajo la formulación (2): “No es posible que la nada produzca algo”. Como antes se señalaba (2) es otra forma de expresar el principio de causalidad cuando en su lectura se lo supone

⁴ No es otra la respuesta de Hume. [Hume (1739-40), B.I, p.iii, sec. iii (1978, p. 81)]. igualmente se debe a Hume la atribución a Locke del argumento anterior. No obstante, es pausable que en el párrafo citado Locke no pretendiera tanto la demostración del principio de causalidad, como la demostración de la existencia de Dios desde su *asunción*.

⁵ No se trata con esto de afirmar que “No es posible que “nada” cause “algo” significa que “algo” tiene que ser causado por “algo”, sino que sólo si así se entiende podemos aceptar el argumento de Locke y que, en cualquier caso, así parece haberlo entendido él.

⁶ Descartes (1647), med. 3 [1977, pp. 35-6]

equivalente a “todo (lo que es) tiene que ser producido por algo”; que así lo entiende Descartes es claro en su conclusión sobre la existencia de la piedra.

Para afirmar la verdad de (2) Descartes parte de que (1): “La realidad del efecto está contenida en la causa” es algo obvio. El razonamiento cartesiano para el principio de causalidad es entonces que, *puesto que la realidad del efecto está contenida en la realidad de su causa, entonces toda realidad proviene de algo*.

[Su argumento para demostrar la existencia de Dios seguirá: En tanto que mi idea (la de Descartes) de Dios contiene más realidad que ninguna otra, entonces Dios es la causa de la idea de Dios y, por tanto, Dios existe.]

No obstante, no es claro cómo puede aceptarse su inferencia a menos que se asuma que todo lo que tiene realidad es causado. Es decir, a no ser que se asuma que todo lo que tiene realidad es *efecto*, del hecho de que todo lo que es efecto debe su existencia (o realidad) a su causa, no se sigue que todo lo que tiene realidad sea causado. De nuevo parece, por tanto, que se presupone lo que se pretende demostrar.

(1) La realidad del efecto está contenida en la causa.

(P) *Todo lo que tiene realidad es efecto.*

 (2) No es posible que la nada produzca algo real. <---> *Algo real tiene que ser producido por algo real.*

4. Hobbes: Todo comienzo de existencia tiene una causa

[A] man cannot imagine anything to begin to exist without a cause, can no other way be made known but by trying how he can imagine it. But if he try, he shall find as much reason, if there be no cause of the thing, to conceive it should begin at one time as at another, that is, he has equal reason to think it should begin at all times; which is impossible, and therefore he must think there was some special cause why it began then rather than sooner or later; or else that it began never, but was eternal⁷.

Hobbes escoge como formulación del principio de causalidad (3): “Todo lo que comienza a existir debe ser causado” para excluir de lo que es necesariamente causado a Dios (que es eterno y, consecuentemente, su existencia no tiene comienzo). Su razonamiento a favor del principio parece ser, entonces, que si [(1)] todo lo que comienza a existir lo hace en un momento dado, y si [(2)] hay una causa por la que algo que comienza a existir comienza en ese momento y no en otro, entonces es claro que [(3)] todo lo que

⁷ Hobbes (1654), prop. 33 [1999, pp. 39-40.]

comienza a existir debe ser causado (o si no, no comenzar a existir, o existir desde siempre).

La aceptación del argumento de Hobbes depende de la aceptación de sus premisas, y, en concreto de (2) -podemos asumir (1). Ahora bien, no parece existir otra razón para la aceptación de (2) que la verdad del propio principio de causalidad. Puesto que, si no es absurdo suponer que algo pueda ocurrir sin una causa, dicha suposición no se convierte en absurda cuando en ella se añade que lo incausado (i.e. algo que empieza a existir) ocurre en un tiempo. Que si algo empieza a existir ahora, y no antes, debe tener una causa por la que empieza a existir en ese tiempo, no tiene más evidencia en sí que la que supone que todo comienzo de existencia es causado. No es lícito, por tanto, intentar demostrar la segunda afirmación desde la primera⁸.

(1) Lo que empieza a existir lo hace en un momento dado.

(2) *Lo que empieza a existir debe tener una causa.* ---> Lo que empieza a existir en un momento dado debe tener una causa de que empiece a existir en ese momento y no, por ejemplo, antes.

(3) Si algo empieza a existir debe tener una causa

5. Ducasse: el efecto no sería el mismo sin su causa

Un intento más contemporáneo para defender el apriorístico conocimiento del principio de causalidad es el de Ducasse:

(...) *a priori* we may go further and prove that the particular hammer blow was *necessary* as well as sufficient to the effect taken in all its full concreteness. (...) The supposition that the identical vase could have broken as it did when it did in the absence of the identical hammer blow, thus involves a contradiction, and we may therefore conclude that that hammer blow, such as it was in all its concreteness, if it was sufficient to the full concrete effect, was also necessary to it.

It is worth while to stress the point which the above argument incidentally brings out, namely that the *universality of causation*, namely, the assertion, 'Every change or state has a cause' or 'No change occurs and no state persists without a cause' *is known to be true merely as a corollary of the law of identity* when the word cause

⁸ De nuevo, ésta parece ser la respuesta de Hume a Hobbes. [Hume (1739-40), B.I, p.iii, sec. iii (1978, p. 80)]

is once properly defined, provided only that *anything occurs at all*⁹.

Ducasse pretende demostrar la verdad del principio de causalidad desde el principio de no contradicción y la necesidad que entraña el vínculo causal. Argumenta que [(1)] si una ocurrencia dada tuviese una causa diferente, entonces sería otra ocurrencia, lo que es imposible (dado el principio de no contradicción). Por tanto, [(2)] toda ocurrencia es necesariamente causada por su causa, y esto, obviamente, entraña que [(3)] toda ocurrencia es necesariamente causada. *Es imposible que una ocurrencia dada tenga otra causa; por tanto, una ocurrencia dada tiene necesariamente su (y, consecuentemente, una) causa.*

(1) es intencionalmente ambiguo. Pero su ambigüedad se hace necesaria para mostrar el razonamiento de Ducasse. Lo que en realidad se sigue de la consideración del ejemplo de la cita anterior es que si una ocurrencia, que de hecho es efecto, tuviese una causa diferente, entonces sería otra ocurrencia (lo que es contradictorio): si el romperse del jarrón hubiese tenido otra causa diferente al golpe de martillo que efectivamente lo causó, no habría sido el mismo romperse del jarrón. Y esto es contradictorio. Ahora bien, de aquí no se sigue (2), sino que necesariamente, toda ocurrencia, que es causada, es causada por su causa. Y, claramente de aquí no se sigue (3) a menos que se asuma que toda ocurrencia es efecto, i.e., a menos que se asuma (3) mismo. Lo más parecido a (3) que se sigue es que toda ocurrencia, que es causada, es causada. Pero esto es una tautología que en poco, o en nada, valida el principio de causalidad.

(1) Si una ocurrencia, *que es causada, i.e. que es efecto* tuviese una causa diferente, entonces sería otra ocurrencia *causada, i.e. otro efecto* (lo que es contradictorio).

(2) Necesariamente, toda ocurrencia *que es causada, i.e. que es efecto* es causada por su causa (*i.e. la causalidad entraña una conexión necesaria*).

(3) Toda ocurrencia *que es causada* es causada.

6. Sullivan: todo, excepto el ser necesario, tiene una causa

En los últimos años, los filósofos analíticos vienen demostrando un interés más bien nulo por el principio de causalidad. Una excepción la constituye Sullivan (1990), (1994). Sullivan defiende que el principio de causalidad es cierto porque no hay razones para creer que es falso, y sólo hay

⁹ Ducasse (1924), pp. 63-64. Cursivas en el original.

razones para creer que es verdadero, si aceptamos que [(3)] al menos algo tiene una causa, que [(2)] la causalidad entraña la necesaria ocurrencia de la causa dado el efecto, y que todo lo que ocurre es contingente (i.e. que es tal que podría no haber ocurrido). Su argumento tiene dos partes. La primera es la siguiente reducción al absurdo:

- (1) Para todo x , x puede comenzar a existir sin una causa.
- (2) Para todo x , si x tiene una causa, entonces x no puede comenzar a existir sin esa causa.
- (3) Al menos algo a tiene una causa b .

- (4) Existe al menos un x tal que x no puede comenzar a existir sin una causa¹⁰.

En la primera parte de su argumento, Sullivan muestra que es falso que (1): “todo es tal que puede empezar sin una causa”. Ahora bien, ni (1) es la negación del principio de causalidad, ni la negación del principio de causalidad supone (1), ni (4) es el principio de causalidad, ni (4) supone el principio de causalidad:

- (4) Existe al menos un x tal que x no puede comenzar a existir sin una causa.
- (PC) Para todo x , si x comienza a existir, entonces x tiene una causa.
- (NPC) Existe un x tal que x comienza a existir sin una causa.
- (1) Para todo x , x puede comenzar a existir sin una causa.

Del hecho de que algo, que si tiene una causa para comenzar a existir, la tiene necesariamente, no se sigue que todo lo que comienza a existir tenga necesariamente una causa. Por tanto, aunque sea imposible que un algo dado empiece a existir sin una causa, esto no indica que todo haya de tener una causa para comenzar a existir.

Sullivan, por su parte, no quiere demostrar con (1)-(4) que el principio de causalidad es cierto, sino más bien, que no hay razones para creer que sea falso y sí razones para creer que es cierto. No obstante, no está claro por qué el que exista algo que para comenzar a ser necesariamente tiene una causa, muestra que no hay razones para creer en la falsedad del principio de causalidad. En esta afirmación no hay nada por lo que negar la posibilidad de que todo, menos algo (que no existiría sino por su causa) sea incausado. Aceptar que hay un cisne que, por algún absurdo, está necesariamente en el lago no es razón para poner entre paréntesis la afirmación de que no todos los cisnes tienen que estar en el lago. Para ello, es preciso demostrar primero que ser cisne y estar en el lago tienen alguna relación de dependencia. Un cisne

¹⁰ Sullivan (1990), pp 267-8 y Sullivan (1994), p. 329.

podría estar necesariamente en el lago porque es cierto, por seguir el paralelismo forzado, que cualquier cosa que está en el lago está necesariamente en el lago.

Se hace necesario pues atender a la segunda parte del argumento de Sullivan:

If I am not invalidly inferring “Nothing can” from “One thing cannot”, how do I get from the one point to the other? By arguing that if we believe that at least one contingent entity is such that necessarily its coming to be has a cause, then we have no reason not to believe this is true of all contingent entities that come to be. For all contingent entities agree with respect to the relevant property - being a contingent entity¹¹.

Si puede demostrarse, parece decir Sullivan, que (4) “existe un x tal que x no puede comenzar a existir sin un causa”, entonces no hay razón para no creer (y sí estamos justificados en creer) que “para todo x , si x comienza a existir x tiene necesariamente una causa” y, por tanto, podemos aceptar que el principio de causalidad es cierto.

La certeza de la primera parte del condicional viene dada por la reducción al absurdo anterior. Pero, ¿cómo se sigue desde aquí el consecuente?. El salto lo justifica Sullivan desde la idea de que todo lo que ocurre o comienza a existir es contingente:

Any need for things to be caused does not seem to depend on a thing’s nature but on the essential structure of a thing *qua* thing; if any thing that comes to be essentially has a cause, then each thing that comes to be essentially has a cause¹².

La referencia a la estructura esencial de una cosa *qua* cosa, es una llamada a la consideración del carácter esencial de toda cosa: su contingencia. Este argumento es, en un sentido fundamental, paralelo al argumento cosmológico, en la versión de Tomás de Aquino, para probar que todo, excepto la primera causa (i.e. Dios), es causado:

Aunque la relación a la causa no entre en la definición del ente que es causado, esta relación es, sin embargo, una consecuencia necesaria de la esencia de lo causado; porque, de ser ente por participación, se sigue que ha de ser causado por otro. Por consiguiente, tal ser no puede existir sin ser causado, al modo como tampoco puede existir el hombre sin ser visible. Sin embargo, como el ser causado no

¹¹ Sullivan (1994), p. 330.

¹² Ibid., citando y aceptando las palabras de Quentin Smith (1994), p. 318.

pertenece en modo alguno a la esencia del ente en cuanto tal, por eso existe un ente no causado¹³.

Tomás de Aquino está haciendo aquí dos afirmaciones fundamentales, paralelas en importantes aspectos a las ideas que Sullivan viene defendiendo. Por un lado, está afirmando que es obvio que en la definición o esencia de algo que es causado no está el que sea causado. Aceptemos por hipótesis, aunque es dudoso que él hubiese utilizado semejante ejemplo, que el movimiento de una bola de billar causa el movimiento de otra. Podemos interpretar a Tomás de Aquino diciendo que no hay nada en el movimiento de la segunda bola, nada en su esencia o definición, que haga referencia al movimiento de la primera ni señale relación alguna con él como causa.

En un importante sentido, Aquino está adelantando el argumento humeano en contra del conocimiento *a priori* de una conexión necesaria entre causa y efecto, o de conocimiento *a priori* del principio de causalidad. Contra el segundo, escribía Hume:

[I]t will be easy for us to conceive any object to be non-existent this moment, and existent the next, without conjoining to it the distinct idea of a cause or productive principle. The separation, therefore, of the idea of a cause from that of a beginning of existence, is plainly possible for the imagination; and consequently the actual separation of these objects is so far possible, that it implies no contradiction nor absurdity; and is therefore incapable of being refuted by any reasoning from mere ideas¹⁴.

El argumento de Hume contra el conocimiento *a priori* del principio de causalidad es bastante semejante a la primera afirmación de Aquino aplicado a la idea, concepto o definición de “comienzo de existencia de algo”. La idea, dice Hume, de “comienzo de existencia de algo” es claramente separable de la idea de “causa”. Por tanto, no podemos argumentar que todo es causado desde el hecho de que todo comienza a existir.

En segundo lugar, y ésta es la afirmación más relevante al caso, Aquino defiende que, en tanto todo -lo que ocurre en el mundo- es tal que podría no haber ocurrido (o en tanto todo es esencialmente contingente, o en tanto todo es tal que el ser no le pertenece esencialmente), entonces todo lo que ocurre ocurre por otro (o debe su ser a otro).

Si eliminamos de esta argumentación (i) la asunción que en él parece implícita acerca de la predicación de la existencia y (ii) lo que podría ser una referencia a la doctrina platónica de la causalidad por participación, el

¹³ Aquino *Summa Theologica*, I, q. 44, a. 1 (1959, p. 498).

¹⁴ Hume (1739-40), B.I, p.iii, sec. iii (1978, p. 80).

razonamiento puede reconstruirse del siguiente modo: *puesto que todo lo que existe en el mundo es tal que podría no haber ocurrido, entonces todo lo que existe en el mundo debe su ocurrir a otro.*

Pero esta razón no es convincente: desde la (necesariamente) posible no ocurrencia de algo que ocurre (i.e. desde la idea de que todo es tal que podría no haber ocurrido) se sigue tanto que hay una causa para su ocurrir como que no hay una causa para su ocurrir. Contingencia y causalidad son conceptos, en principio, separables.

Del mismo modo, por tanto, es posible argumentar contra Sullivan. Del hecho de que sepamos que algo que empieza existir es (necesariamente) contingente y *además* necesariamente causado, no se sigue que hayamos de establecer alguna relación entre su contingencia y su ser causado y, por tanto, no se sigue todo lo que es (esencialmente) contingente haya de ser (esencialmente) causado, tal y como defiende Sullivan.¹⁵ Supongamos que tenemos el conjunto de todas las cosas (necesariamente) contingentes, parece que aún podemos decir que algunas empiezan a existir (necesariamente) causadas y otras no, i.e. que otras comienzan a existir por azar o sin una causa.

Contra esta idea escribe Sullivan:

It would be entirely arbitrary to say that a contingent entity needs a cause for its emergence provided that it is blue, but not if it is red. The relevant property is not its colour or its size, but its contingency¹⁶.

Sullivan parece estar diciendo que no es sino arbitrario dividir el conjunto de las cosas contingentes que empiezan a existir en función de *cualquier* propiedad y afirmar, en consecuencia, que unas son (necesariamente) causadas y otras pueden no serlo.

Pero a esto puede igualmente ser replicado que no hace falta dividir el conjunto de las cosas contingentes en función de *alguna* propiedad para, en

¹⁵ Se podría pensar que en la crítica a Sullivan hay una premisa que no está siendo considerada en toda su relevancia: la idea de que todo lo que tiene una causa tiene necesariamente una causa (i.e. la idea de que la relación causal entraña una conexión necesaria en algún sentido). Que esta idea y el principio de causalidad son independientes, sin embargo, se ha considerado al analizar el argumento de Ducasse (ver más arriba). Que algo es contingente y que algo tiene una causa, son aplicables separadamente incluso aunque la causalidad entraña una conexión necesaria. Que algo ocurre aunque podría no haber ocurrido conlleva tanto que necesariamente tiene una causa (si es causado), como que no es causado. Igualmente, aún si la causalidad entraña algún tipo de conexión necesaria, algo puede ser causado es compatible y contingencia. Si la causa supone la necesaria ocurrencia del efecto, entonces el efecto aún podría no haber sucedido: si su causa no hubiese sucedido. Si el efecto supone la necesaria ocurrencia de la causa, su causa aún podría no haber sucedido: en caso de que el efecto no hubiese sucedido; i.e., el efecto aún podría no haber ocurrido.

¹⁶ Sullivan (1994), p. 330.

consecuencia, determinar que unas son causadas y otras no. Basta dividir el conjunto de las cosas contingentes en función de si son causadas o no. No es necesario buscar ninguna propiedad, que algunas ocurrencias tienen y otras no, y así mostrar cómo unas son necesariamente causadas y otras no: la propiedad que sirve para su clasificación es no más que la de ser causadas, o no. Y, de nuevo, mientras no se muestre alguna dependencia directa entre “ser contingente” (o “ser esencialmente contingente”) y “ser causado” (o “ser necesariamente causado”), es tan arbitrario negar la división como afirmarla.

7. Quentin Smith: todo excepto el comienzo del universo tiene un origen

Quentin Smith (1994) defiende que aunque, tal y como Sullivan demuestra desde el supuesto de que algo es causado, es falso que todo es tal que puede comenzar a existir sin una causa, y aunque es cierto que todo lo que comienza a existir necesariamente tiene una causa, aún puede defenderse que el comienzo de la existencia del universo no tiene una causa. Y esto es así porque “todo” en los supuestos anteriores, significa “todo lo que pertenece a historia del universo”. Por tanto, puesto que el comienzo de la existencia del universo no pertenece a la historia del universo, sino que es el principio de su propia historia, entonces aún es posible que el comienzo del universo sea la excepción a la regla “todo lo que empieza a ser necesariamente tiene una causa”:

One thing that comes to be can come to be without a cause, but each other thing that comes to be necessarily has a cause of its coming to be¹⁷.

Aquí no nos interesan tanto las razones de Quentin Smith para defender que el comienzo del universo es incausado, como su defensa del principio de causalidad bajo la fórmula, “todo lo que comienza a existir, con la excepción del comienzo del universo mismo, necesariamente tiene una causa”. La defensa de esta idea la realiza Smith simplemente apelando al principio kripkeano de la necesidad de origen:

The theory of the necessity of origins is that it is really (metaphysically) impossible that something come to be but not from the origins it in fact has. Given the theory of the necessity of origins (...) it follows that it is metaphysically impossible that each thing come to be without a cause¹⁸.

Seguramente un pasaje relevante en el que Kripke defiende el principio de la necesidad de origen es el siguiente:

¹⁷ Smith (1994), p. 318.

¹⁸ Ibid., p. 315.

A principle suggested (...) is: *If a material object has its origin from a certain hunk of matter, it could not have had its origin in any other matter.* (...) Let 'B' be a name (rigid designator) of a table, let 'A' name the piece of wood from which it actually came. Let 'C' name another piece of wood. Then suppose B were made from A, as in the actual world, but also another table C were simultaneously made from C. Now in this situation $B \neq D$; hence, even if D were made by itself, and no table were made from A, D would not be B¹⁹.

Traer al caso la teoría de Kripke nos mete en un montón de consideraciones y problemas que, hasta el momento, han sido innecesarias o inexistentes. De ellos, el más relevante concierne a la naturaleza de los elementos causalmente relacionados. De momento, nada se ha dicho al respecto, y los argumentos hasta ahora presentados son igualmente aplicables a hechos, sucesos, propiedades universales, propiedades particulares, o individuos. En cualquier caso, si la pregunta es ¿demuestra el principio de la necesidad de origen, el principio de causalidad (para todo, menos para el origen del universo)?, la respuesta parece ser que no:

(i) En primer lugar, en tanto el principio de la necesidad de origen se refiere a entidades (materiales) con nombres propios que no esconden descripciones definidas (i.e. entidades a las que referimos directamente; i.e. entidades que son "algo" más que sumas de propiedades), es necesario argumentar primero la existencia de tales entidades. Pero, incluso aunque se garantice su existencia,

(ii) es necesario argumentar que el principio de causalidad hace referencia *exclusivamente* a estas entidades. Es decir, habría que defender: (ii.i) que sólo estas entidades son causadas: apelar al principio de Kripke para argumentar, por ejemplo, que las propiedades son causadas, es inútil; habría que mostrar por qué las propiedades y demás ocurrencias no son causadas. Pero, entonces, habría que defender también que (ii.ii) todo lo demás que parece ocurrir, no tiene "realidad", i.e., es tal que el hecho de que no sea causado no falsea el principio de causalidad. Sin embargo,

(iii) ni siquiera garantizando (i) y (ii), se sigue, desde el principio de la necesidad de origen, el principio de causalidad. Si se quiere defender éste desde el primero aún hay que argumentar que el origen necesario de Kripke es un origen "que origina" o que causa, y que no es sólo "origen" al modo en que una mesa es "de" madera. Puesto que (iii.i) si la necesidad de origen lo que supone es que toda entidad material es necesariamente "de" la materia que es, lo más que significaría el principio de causalidad es que toda entidad material es "de"

¹⁹ Kripke (1980), p. 114, n. 56. Cursivas en el original.

alguna materia (i.e. sería un principio vacío). Pero (iii.ii) si el origen kripkeano puede equipararse de alguna manera a algún principio causal eficiente (tal vez, al modo en que los padres “originan” a sus hijos), entonces a no ser que *presupongamos* que todas las entidades son originadas por algo (i.e., a no ser que presupongamos el principio de causalidad mismo -con todas las anteriores consideraciones), no puede concluirse desde el principio de Kripke que “todo, menos el comienzo del universo, tiene necesariamente un origen”.

Conclusiones

Hasta aquí he presentado algunos (ni todos los que son²⁰, ni todos los que podrían ser) argumentos que trataban de demostrar que el principio de causalidad es cierto. Su consideración muestra, al menos, las grandes dificultades que supone montar una defensa *a priori* del mismo. Pero quizás algo más puede concluirse de este análisis: la necesidad de demostrar, para ello, la existencia de algún vínculo fundamental entre lo que es suceder y lo que es ser causado. Añadir que todo lo que ocurre es (incluso, por definición) temporal, o contingente, material, que tiene un comienzo de existencia o considerar la definición o el concepto de lo que sea que concretamente ocurre, son insuficientes (o innecesarias) restricciones para mostrar este vínculo. Por otro lado, defender que la causalidad entraña una conexión necesaria o que hay ocurrencias que son, de hecho, efectos, tampoco facilita la labor. Aún así, como reconocía el propio Hume, todavía es posible que el principio de causalidad sea cierto o hasta necesariamente cierto. La dudoso es, no obstante, que sólo el análisis filosófico pueda establecer su valor de verdad[†].

BIBLIOGRAFÍA

- Aquino, Tomás de *Summa Theologica* [Traduc. (1959) Madrid: B.A.C.]
Bunge, Mario Augusto (1961) *Causalidad: El Principio de Causalidad en la Ciencia Moderna* Buenos Aires: Editorial Universitaria de Buenos Aires (1ª ed. 1959).
Descartes, René (1647) *Meditaciones metafísicas con objeciones y respuestas* [Traduc. (1977) Madrid: Alfaguara]
Ducasse, John Curt (1924) *Causation and the Types of Necessity* University of Washington, Publications in the Social Sciences.
Hobbes, Thomas (1654) *Treatise 'Of Liberty and Necessity'*. En Vere Chappell (ed.) (1999) *Hobbes and Bramhall on Liberty and Necessity* Cambridge: Cambridge University Press, pp. 15-42.
Hume, David (1739-40) *A Treatise of Human Nature* [Ed. (1978) Oxford: Clarendon Press]
Kripke, Saul (1980) *Naming and Necessity* Cambridge: Harvard University Press.
Locke, John (1690) *Ensayo sobre el entendimiento humano* [Traduc. Méjico: Fondo de Cultura Económica]

²⁰ Grandes ausencias son, al menos, el argumento kantiano y el argumento de Spinoza.

[†] Este trabajo fue finalizado gracias a una Beca de Formación de Investigadores del Dpto. de Educación del Gobierno Vasco.

Smith, Quentin (1994) "Can Everything Come to Be Without a Cause?" *Dialogue*, 33: 313-323.

Sullivan, Thomas D. (1990) "Coming to be without a Cause". *Philosophy*, 65: 261-270.

Sullivan, Thomas D. (1994) "On the Alleged Causeless Beginning of the Universe: A Reply to Quentin Smith" *Dialogue*, 33: 325-335.

Maria José García Encinas

Rutgers University

e-mail: ylbgaenm@sf.ehu.es



REALISMO Y RELATIVISMO EN LA FILOSOFÍA DE PUTNAM*

JOSÉ ANTONIO GUERRERO DEL AMO
Universidad Complutense de Madrid

ABSTRACT: This paper discusses some problems of Putnam internal realism, that made it something unsustainable. These would have made him change his mind and propose a natural realism, that denies the existence of intermediaries between the perceiver and the object perceived, and that has some proximity with metaphysical realism. This natural realism, apart from inhering many of the problems of internal realism, poses some of its own, due, in part, to the vagueness of its formulation. An attempt to state it more explicitly poses not only more problems, but also brings it nearer to the metaphysical realism Putnam has been criticizing.

KEYWORDS: Internal realism, natural realism, metaphysical realism.

I

El objetivo de este trabajo es plantear y discutir algunos de los problemas que, a mi juicio, presentan tanto el realismo interno o pragmático como el más reciente realismo natural de Putnam¹. Algunos de esos problemas del realismo interno lo habrían convertido en algo insostenible, de manera que, en sus últimos escritos, los publicados en los últimos 10 años aproximadamente, Putnam ha cambiado algunos de sus puntos de vista y defiende una forma de realismo que el gusta de calificar como realismo natural (1992b, 356; 1994c, 454) y que considera próxima al realismo del sentido común. Esta denominación no oculta, sin embargo, que es una forma de realismo que guarda un cierto parentesco con el realismo metafísico que ha criticado durante mucho tiempo. El realismo natural, en la medida en que todavía mantiene bastantes componentes del realismo interno, también heredaría muchos de sus problemas, a los que habría que añadir los que

* Este trabajo se ha realizado mientras se disfrutaba de la ayuda financiera del proyecto de investigación PB96-0580/96 de la DGICYT.

¹ El que este trabajo se centre en los aspectos problemáticos de la posición realista de Putnam, no significa que no valoremos muy positivamente otros componentes de su propuesta.

surgen cuando una formulación tan genérica como la que hace Putnam se trata de concretar, para que pueda responder a las cuestiones que pretende solucionar. Esa concreción del mismo, además, aumenta su proximidad al realismo metafísico.

II

El problema que, según cuenta el propio Putnam (1994c, 459), le habría llevado a formular el realismo interno, era que no sabía cómo defender el realismo, pero tampoco veía ninguna otra forma de entender la relación del lenguaje con la realidad. Como todos ustedes conocen, la solución que dio a este problema, en esa etapa de su pensamiento, consistió en una forma de realismo, que más bien era un antirrealismo, basado en lo que se ha denominado una semántica verificacionista. Esa forma de realismo sostenía que “los propios ‘objetos’ son tanto construidos como descubiertos, son tanto producto de nuestra invención conceptual como del factor objetivo de la experiencia” (1981/1988, 63), no teniendo, por tanto, sentido hablar de ellos independientemente de nuestros esquemas conceptuales, pero tampoco considerarlos como productos de ellos. La semántica verificacionista, inspirada en M. Dummett, defendía que “en la perspectiva internalista, la verdad es una especie de aceptabilidad racional (idealizada) –una especie de coherencia ideal de nuestras creencias entre sí y con nuestras experiencias, considerándolas como experiencias representadas en nuestro sistema de creencias- y no una correspondencia con “estados de cosas” independientes de la mente o del discurso. No existe un punto de vista como el del Ojo Divino que podamos conocer o imaginar con provecho” (Putnam, 1981/1988, 59; cfr. también 1982, 41; 1986/1990, 115).

Los problemas que presenta una teoría semejante son varios. En primer lugar, si se identifica la verdad con aceptabilidad justificada en condiciones ideales, cabe plantearse, como lo hace en algún momento el propio Putnam (1992b/ 1999, 42 n. 26), ¿cuándo son ideales las condiciones epistémicas? La respuesta de nuestro pensador es que “lo son cuando son suficientemente buenas para determinar si la afirmación en cuestión es verdadera o falsa”. Evidentemente esta respuesta es un círculo vicioso y nos deja en el mismo lugar que estábamos. Putnam admite que considerándola como una definición de verdad es así, pero señala que él no la considera de ese modo, sino que cree que la noción de verdad no puede ser reducida a conceptos que no la presupongan de una forma u otra. Y añade: “mi idea es negativa: no concebimos la verdad (o, en lo que aquí respecta las buenas condiciones de verdad) como algo que sobrepase totalmente el *uso*” (*ibid.*). Entender nuestro lenguaje, pensaba Putnam entonces, consiste en el dominio de su uso (1992d, 282; cfr. también 1977/1994, 38), y, además, creía que “los diversos usos de las palabras podían ser claramente organizados y representados por algo semejante a un programa informático” (1992a/1994,

231). Recuérdese que Putnam todavía defendía el funcionalismo. Ahora bien, ¿no nos deja esto en una situación tan mala y peor que la verdad como correspondencia del realismo metafísico que tanto ha criticado? Si no se pueden establecer cuáles son las condiciones ideales en las que podemos saber si una teoría es verdadera o no, este concepto de verdad resulta tan misterioso como la referencia en el concepto de verdad como correspondencia.

Pero, no sólo eso. El realismo interno tiene que afrontar una cuestión todavía más grave y ésta es que no resuelve el problema que Putnam pretendía solucionar con él, el problema de cómo entender la relación entre lenguaje y realidad, el problema de cómo podemos tener acceso referencial en situaciones epistémicas suficientemente buenas. No sólo es que sea problemático, como acabamos de decir, cuáles sean esas situaciones suficientemente buenas, sino que se está dando por supuesto lo que es una situación epistémica. Y la concepción que el Putnam realista interno tenía de una situación epistémica era, en aspectos importantes, semejante a la tradicional, en el sentido de que postulaba la existencia de un intermediario entre la realidad y el sujeto cognoscente (1977/1994, 27). Ahora bien, si se admite la existencia de ese intermediario, sigue siendo un problema el modo cómo se accede a la realidad desde ese intermediario. El problema es el mismo que Putnam ha venido criticando al realismo metafísico.

Por otra parte, si lo que es verdadero se establece por el uso, ¿no nos aboca esto irremediabilmente al relativismo? Esta perspectiva internalista que venimos discutiendo implicaba, según Putnam, que “hay más de una teoría o descripción del mundo verdadera” (1981/1988, 59). Dicha tesis se debía a la relatividad conceptual. Esta consistiría en lo opuesto al punto de vista del ojo de Dios, esto es, en la aceptación de que existen distintas formas de representar y describir la realidad y no hay ninguna de ellas que se pueda considerar *la* correcta. Esto llevaba, entre otras cosas, a que teorías incompatibles, pero intertraducibles, pueden ser las dos verdaderas al mismo tiempo. Ahora bien, si diferentes esquemas conceptuales pueden ser igualmente correctos y lo único que ayuda a decidir entre ellos no son nada más que sus cualidades epistémicas, parece inevitable concluir que la objetividad se ha perdido para siempre y que el relativismo de nuestros esquemas conceptuales es la única posibilidad.

Ante esto, Putnam ha respondido, por una parte, que la relatividad conceptual no implica relativismo (1987a/1994, 61 y ss.) y, por otra, que, aunque es cierto que puede haber en un momento dado varias descripciones alternativas del mundo que sean correctas (verdaderas), no por eso todas esas descripciones son igual de correctas o, si se prefiere, igual de válidas (Putnam, 1986/1990, 114).

La razón que ha esgrimido para defender que la relatividad conceptual no implicaba relativismo es que una vez que se ha declarado dentro de qué marco conceptual nos movemos, la respuesta a la pregunta acerca de lo que

existe ya no es, en absoluto, asunto de convención (1987/1994, 64). Esto implica que no perderíamos totalmente la objetividad. Pero el tipo de objetividad que tendríamos sería relativa a nuestros esquemas conceptuales. Posiblemente, como dice Putnam, es la única objetividad a que podemos aspirar. Ahora bien, no tiene mucho sentido negar que esto es relativismo, aunque no lo sea de un tipo extremo.

Las razones de Putnam para defender que, entre descripciones alternativas verdaderas, hay unas mejores que otras son mucho más difusas. Poder afirmar eso exigiría disponer de unos criterios para comparar y valorar esquemas conceptuales que dan origen a descripciones alternativas y, en su caso, elegir uno. Dichos criterios, además, deberían estar fuera de cualquier marco conceptual, si no queremos caer en alguna forma de relativismo. Hay un texto en "Why is a Philosopher?" que, además de intentar responder estas cuestiones, plantea en toda su crudeza la problemática que acabamos de exponer. El texto dice lo siguiente:

Supuesto que reconocemos que hay mejores y peores textos (teorías), 'mejor' y 'peor' pueden depender de nuestra situación histórica y de nuestros propósitos; no hay aquí ninguna noción de un Punto de Vista del Ojo de Dios de la verdad. Pero la noción de una correcta (o al menos una 'mejor') respuesta a una cuestión está sujeta a dos constricciones: 1) la corrección no es subjetiva. (...) 2) La corrección va más allá de la justificación. (...) Mi punto de vista es que la verdad tiene que ser identificada con justificación idealizada, más bien que con justificación en evidencia presente. "Verdad" en este sentido es tan sensible al contexto como lo somos nosotros. Las condiciones de asertabilidad para un enunciado arbitrario no son especificables. Si las condiciones de asertabilidad no son especificables, ¿cómo las aprendemos? Las aprendemos adquiriendo una práctica... [Pero] lo que adquirimos no es un conocimiento que pueda ser aplicado como si fuera un algoritmo" (1986/1990, 114-5).

Quizás lo más significativo de este texto es la lucha de Putnam por escapar al relativismo, sin poder lograrlo. Por una parte, admite que nuestras valoraciones dependen del momento histórico y de los propósitos, es decir, son relativas; por otra, trata de evitar el relativismo señalando que la corrección no es subjetiva y va más allá de la justificación. Ahora bien, sería conveniente que Putnam nos explicara en qué consiste ese "no ser subjetivo" y ese "ir más allá de la justificación", si sus palabras quieren ser algo más que la expresión de un deseo. Hay demasiada frase negativa y poca propuesta afirmativa en todo esto y más bien parece un intento vano de escapar al relativismo, un tipo de relativismo, por cierto, que no está muy lejos del relativismo cultural que Putnam atribuye a Rorty y por el que tanto le critica (1993/1994b, 297-300; 1997, 4-5).

Por lo demás, estos puntos de vista los sigue manteniendo Putnam en la actualidad (1997, 12), a pesar de haber adoptado un realismo natural. En consecuencia, el realismo natural heredará todos estos problemas.

Otros intentos de Putnam por enfrentarse a estas cuestiones creo que terminan en la misma frustración. Así, en distintos lugares, ha sugerido que la validez de los esquemas conceptuales derivaría de las necesidades (1987a/1994, 151) y los intereses humanos (1992a/1994, 108). Ahora bien, en el caso de las primeras, ¿cómo establecemos cuáles son esas necesidades, independientemente de nuestros esquemas conceptuales? No parece que eso sea posible, y la respuesta de Putnam es cualquier cosa menos satisfactoria: “Es porque hay necesidades humanas reales, y no simplemente deseos, por lo que tiene sentido distinguir entre valores mejores y peores, y, por lo mismo, entre cuchillos mejores y peores. ¿Qué son esas necesidades humanas pre-existentes y cómo se distinguen de los meros deseos? Aquí, Dewey, como Goodman, nos dice que las necesidades humanas tampoco pre-existen, que la humanidad está constantemente rediseñándose a sí misma, y que nosotros creamos las necesidades. (...) Nuestras nociones –la noción de un valor, la noción de una imagen moral, la noción de un criterio, la noción de una necesidad- están tan entrelazadas que ninguna de ellas puede proporcionar un fundamento para la ética” (1987a/1994, 151). Esto, una vez más, nos aboca al relativismo. En cuanto a los intereses, aunque insiste en que “que un interés sea o no relevante es algo que se puede razonar” y en que “decir que una noción es relativa dependiendo del interés no es decir que todos los intereses son igualmente razonables”, cuando se pregunta “¿qué es lo que hace que unos intereses sean más relevantes que otros?”, no le queda más remedio que contestar que “la razonabilidad depende de diferentes cosas en diferentes contextos. No hay una respuesta general” (1992a/1994, 108; cfr. también 1997, 12). Con lo cual volvemos otra vez al punto donde nos encontrábamos antes, esto es, al relativismo. Es cierto que Putnam sigue insistiendo, como veíamos que hacía antes de la corrección, en que, a pesar de que la razonabilidad es relativa a un contexto, incluyendo la cultura, aquella “no es simplemente lo que una cultura considera razonable” (1997, 12). Pero no dice qué es eso que hay que añadir a lo que una cultura considera razonable para obtener la razonabilidad, digamos, objetiva.

En otras ocasiones, insinúa que hay criterios, que son el resultado de la búsqueda de mejores marcos conceptuales, que nos permitirían valorar unos como mejores que otros (1992c/1999, 44, n. 37). Pero estos criterios no parece que se puedan situar fuera de un marco conceptual con lo que volvemos al relativismo de los esquemas conceptuales.

III

Sin duda, ha sido el tomar conciencia de algunos de estos problemas, sobre todo el de la incapacidad del realismo interno para responder a la

cuestión de cómo se conecta el lenguaje con la realidad, dada su aceptación de intermediarios en la percepción, lo que ha llevado a Putnam a abandonar determinados componentes de aquel y proponer un realismo natural. Las causas inmediatas confesadas de ese cambio habrían sido las conferencias John Locke de John McDowell², su larga familiaridad con el pensamiento de Austin y Wittgenstein y su interés durante los últimos 15 años por el trabajo de William James (1992b, 356). Además, a todo esto, creo que hay que añadir otro factor no confesado que yo considero esencial y es el cambio que se ha producido en su pensamiento con respecto a la existencia de una realidad independiente. Como todos ustedes saben, una de las críticas que Putnam ha venido haciendo al realismo metafísico consistía en rechazar su supuesta tesis de que existe una totalidad fija de objetos que existen independientemente de nuestras representaciones. Este rechazo, basado en la relatividad conceptual, lo expresaba señalando que no sabemos de qué estamos hablando cuando sostenemos una tesis semejante (1987/1994, 87). Pero los argumentos de Putnam de ningún modo probaban que no existiera una realidad independiente. Lo máximo que hacían era mostrar la imposibilidad de que se pudiera decir nada de ella fuera de un esquema conceptual. Y, además, su realismo interno suponía la existencia de esa realidad independiente. Si éste consiste, como sostenía, en que “los propios objetos son tanto construidos como descubiertos, son tanto producto de nuestra invención conceptual como del factor ‘objetivo’ de la experiencia, el factor independiente de nuestra voluntad” (1981/1988, 63), ¿no significa esto admitir implícitamente la existencia de una realidad independiente, aunque no podamos decir nada acerca de cómo es independientemente de nuestras representaciones? Y a la misma cuestión se llega, si partimos de otra de las formulaciones más conocida del realismo interno, que señala que “el internalismo no niega que haya *inputs* experienciales en el conocimiento; el conocimiento no es un relato que no tenga otra constricción que la coherencia *interna*; lo que niega es que existan *inputs* que no estén configurados en alguna medida por nuestros conceptos, por el vocabulario que utilizamos para dar cuenta de ellos y para describirlos, o *inputs* que admitan una sola descripción, independiente de toda opción conceptual” (1981/1988, 64). Si admitimos entradas de fuera a la experiencia, estamos admitiendo tácitamente que existe algo independiente de la misma. Por tanto, el realismo interno suponía la tesis de la existencia de una realidad independiente.

Todo esto, con toda probabilidad, ha contribuido a que Putnam aceptara esa tesis. No es que llegue a afirmar que existe una totalidad fija de todos los objetos y todas las propiedades, como, según él, hacía el realista metafísico, pero sus expresiones remiten con toda claridad a una realidad

² Publicadas con el título *Mind and World*, Cambridge (Mass.), Harvard University Press, 1995.

externa independiente. Por ejemplo, en la primera de las conferencias Dewey, admitía que “lo que el realista tradicional está señalando es que cuando yo hablo sobre alguna cosa que no está causalmente afectada por mis propios intereses –digamos, cuando señalo que hay millones de especies de hormigas en el mundo- puedo también decir que el mundo sería igual en este respecto incluso si yo no tuviera esos intereses y no hubiera dado esa descripción, etc. Y con todo esto yo estoy de acuerdo” (1994c, 448, n. 7). Y un poco más adelante vuelve a reconocer que “es verdad que una pretensión de conocimiento tiene que responder ante la realidad, y, en la mayoría de los casos, esto significa una realidad independiente del que habla” (*ibid.*, 449), aunque acto seguido indique que “la reflexión en la experiencia humana sugiere que ni la forma de todas nuestras pretensiones de conocimiento ni los modos en los que ellas son responsables ante la realidad están fijadas de una vez y para siempre, contrariamente a las asunciones del realista tradicional” (*ibid.*).

Dada, pues, la aceptación de esta realidad independiente, era casi inevitable que surgiera con mayor claridad el problema de cómo se conecta esa realidad con el lenguaje, o cómo se conectan nuestros *inputs* con esa realidad externa. Si se piensa que “la mente y el mundo construyen conjuntamente la mente y el mundo” (1981/1988, 13) es mucho más difícil que surja dicha cuestión. Por eso considero este cambio como un factor decisivo a la hora de abandonar en parte el realismo interno y proponer un realismo natural.

Lo que caracteriza este realismo natural, al que Putnam ha llamado con frecuencia realismo directo (1992b, 356; 1992c/1999, 117-118), aunque últimamente prefiere evitar esta denominación (1992d, 10-11; 1994c, 453), es:

1) La admisión de esa realidad independiente de la que venimos hablando.

2) Su preocupación por el tema de la percepción, que considera unido al tema de la conexión entre el lenguaje y la realidad, y la respuesta que da al mismo, a saber, que no hay ningún intermediario entre la realidad exterior y el sujeto cognoscente. Así dice textualmente “como ahora pienso, el hecho de que la percepción no sea infalible no es razón para postular un intermediario (un conjunto de “datos sensibles”, en el cuadro cartesiano tradicional, o una “representación mental” en el de la “ciencia cognitiva” contemporánea) entre la persona que percibe y el objeto percibido)... En suma, lo que espero poder argumentar es que la percepción... es un ejercicio de conceptos, por tanto, no “inocente”, y ciertamente no infalible; pero a la vez, ello no obstante, una forma de contacto *directo* con los objetos externos” (1994a, 44). O como señala en otro lugar, “nosotros podemos *ver* la cosa. (...) Sé perfectamente que la percepción incontaminada no existe, pero esto no significa que ésta no sea una percepción de la forma en que las cosas se encuentran verdaderamente” (1992c/1999, p. 117). Este acceso a los objetos significa no que los

percibamos, en el sentido de tener una relación puramente causal con ellos, sino que tenemos un acceso cognoscitivo a ellos. Ahora, por tanto, la noción de intencionalidad, le resulta totalmente transparente: “La respuesta más simple a la pregunta ‘¿cómo podemos referirnos a un objeto externo, por ejemplo, una silla?’ es: ‘¿Qué problema hay? Nosotros podemos ver la cosa’” (1992c/1999, 117).

3) El mantenimiento de la relatividad conceptual, con todo lo que ésta implicaba de posibilidad de varias teorías o descripciones del mundo verdaderas al mismo tiempo, como uno de sus componentes.

4) Un cambio en el concepto de verdad del realismo interno. Parece que ya no sostiene que la verdad es aceptabilidad justificada en condiciones apropiadas (1987/1994, 15; 1994a, v), al menos no en todos los casos (1997, 7-8), aunque sigue insistiendo en que aquella no se puede establecer reduciéndola a otros conceptos y que, en consecuencia, sigue siendo el *uso* lo que determina la interpretación del lenguaje. Ahora, sin embargo, *uso* hay que entenderlo de un modo completamente diferente a como lo entendía en el realismo interno. Mientras, en éste, la noción de uso “era científico cognitiva, esto es, uso era descrito en gran medida en términos de programas de ordenador en el cerebro”, la noción de uso que ahora propone es semejante a la del último Wittgenstein, a saber, “el uso de las palabras en un juego del lenguaje no puede ser en la mayoría de los casos descrito sin emplear el vocabulario de este juego o un vocabulario relacionado internamente con el vocabulario de este juego. (...) El eslogan realmente no debería ser [ahora] ‘El significado es el uso’, sino ‘Entender es tener las habilidades que uno ejerce cuando usa y al usar el lenguaje’” (1994c, 457-459; cfr. también 1992d, 6). Putnam cree que esas habilidades de ninguna manera se pueden especificar y establecer en forma de algoritmo (1986/1990, 114-5).

5) Finalmente, el realismo natural sostiene que nuestras representaciones tienen que responder ante la realidad (1994c, 449, 452).

Hecha esta breve exposición del realismo natural, vayamos a los problemas. En primer lugar, volver a un realismo directo a finales del siglo XX, aunque sea muy matizado, resulta, cuando menos, un poco sorprendente. Es cierto que, en la medida en que Putnam une a dicho realismo directo la relatividad conceptual, contrarresta muchas de las críticas que se le han hecho al mismo. Pero, en ese caso, no se entiende muy bien qué significa “captar las cosas como verdaderamente son”. Si, por una parte, la ciencia ha puesto de manifiesto que lo que nosotros sentimos depende de nuestro sistema nervioso y de su funcionamiento (incluidos los tipos de receptores que tenemos) y no tanto del tipo de energía que incide en nuestros receptores; y, por otra, aceptamos la relatividad conceptual, ¿en qué sentido conocemos las cosas como son? ¿Qué es tener acceso cognoscitivo a ellas? Una posibilidad, sugerida por un texto del Putnam realista interno, es que sean los propios objetos, que percibimos directamente, los que al ser de una determinada forma

dentro de un esquema conceptual, “intrínsecamente pertenezcan a determinadas etiquetas” de las que forman nuestros esquemas conceptuales³. Esto estaría de acuerdo con la admisión de una realidad independiente que ha hecho y también con la relatividad conceptual que defiende, en el sentido de que los seres humanos podríamos utilizar distintos juegos de etiquetas. La ventaja inmediata de esta propuesta, por tanto, sería el proporcionarnos un concepto de objetividad del que carecía el realismo interno: los objetos tienen una forma de ser intrínseca, que, aunque sólo se manifiesta dentro de los esquemas conceptuales y está condicionada en su manifestación por ellos, no depende sólo de ellos. Esta objetividad aleja a Putnam de un relativismo extremo. Y tendría la virtud de que no implica la fantasía de que “hay una totalidad de ‘formas’ o ‘universales’ o ‘propiedades’ fijas de una vez y para siempre”, ni la existencia de una estructura fija de todos los posibles pensamientos, fijada por las formas (1994c, 449), para no concederle al realista metafísico todo lo que le ha estado negando durante años, a saber, la tesis de una totalidad fija de objetos independientes de la mente y, con ella, la base para otras tesis del realismo metafísico. No estoy del todo seguro, sin embargo, de que Putnam admitiera un tipo de objetividad semejante.

Por otra parte, también presenta problemas el concepto de verdad que va unido al realismo natural. En primer lugar, en la medida en que Putnam siga manteniendo que la verdad es aceptabilidad justificada en condiciones apropiadas (1997, 7-8), las críticas, expuestas anteriormente, referentes a la imposibilidad de establecer esas condiciones ideales siguen siendo válidas.

En segundo lugar, ¿cómo entender que nuestras representaciones tengan que responder ante la realidad? ¿Qué significa aquí “responder”? ¿No se estará insinuando una teoría de la verdad como correspondencia? Por supuesto, sería una correspondencia entre nuestras representaciones y objetos conceptualizados, ya que eso es lo que captamos directamente en la percepción. Además, no habría una sola correspondencia, sino muchas, tantas como esquemas conceptuales. Pero el problema que se plantea entonces es ¿cómo podemos saber cuando hay una correspondencia adecuada, esto es, cuando nuestras representaciones responden ante la realidad, si se rechaza el punto de vista del Ojo de Dios? Parece que la única posibilidad sería que esa corrección en la correspondencia venga determinada por el uso dentro de nuestros esquemas conceptuales.

En tercer lugar, todas las críticas que hacíamos al realismo interno sobre la relatividad que implicaba la noción de uso, se agrandan ahora con el

³ El texto dice: “si como mantengo, los propios ‘objetos’ son tanto construidos como descubiertos, son tanto producto de nuestra invención conceptual como del factor ‘objetivo’ de la experiencia, el factor independiente de nuestra voluntad, entonces los objetos pertenecen intrínsecamente a ciertas etiquetas; porque esas etiquetas son las herramientas que usamos para construir una versión del mundo en la que tales objetos ocupan un lugar preeminente” (1981/1988, 63).

cambio de dicha noción. Ya señalábamos que este problema sigue preocupándole y que sus puntos de vista siguen siendo muy semejantes.

IV

Podemos, por tanto, terminar con las siguientes conclusiones:

1ª) Putnam, en su abandono progresivo del realismo interno y su aceptación de un realismo natural, ha dado pasos importantes hacia un realismo mas cercano al realismo metafísico: ha admitido la existencia de una realidad independiente, ha admitido la posibilidad de conocerla tal y como es (realismo directo) y, por último, ha admitido la necesidad de que las teorías tienen que responder ante la realidad.

2ª) Cuando hemos intentado concretar más su realismo natural, hemos visto que éste lleva a un concepto de objetividad más “objetivo” y a una teoría de la verdad como correspondencia relativa a los esquemas conceptuales, que también lo aproximan al realismo metafísico.

BIBLIOGRAFÍA⁴

- Putnam, H. (1976): “Realismo y razón”, parte iv de Putnam (1978).
 ----- (1977): “Models and Reality”, *Journal of Symbolic Logic*, 45 (1980), pp. 464-482. Trad. cast.: “Los modelos y la realidad”, *Diálogos*, 63 (1994), pp. 7-39.
 ----- (1978): *Meaning and the Moral Sciences*, Londo, Rouledge & Kegan Paul. Trad. cast.: *El significado y las ciencias morales*, México, UNAM, 1991.
 ----- (1981): *Reason, Truth and History*, Cambridge, Cambridge University Press. Trad. cast.: *Razón, verdad e historia*, Madrid, Tecnos, 1988.
 ----- (1982): “A Defense of Internal Realism”, en Putnam (1990), pp. 30-42.
 ----- (1986): “Why Is a Philosopher?”, en Putnam (1990), pp. 105-119.
 ----- (1987a): *The Many Faces of Realism*, La Salle, Illionis, Open Court Publishing Company. Trad. cast.: *Las mil caras del realismo*, Barcelona, Paidós, I. C. E./ U. A. B., 1994.
 ----- (1987b): “Realism with a Human Face”, en Putnam (1990), pp. 3-29.
 ----- (1990): *Realism with a Human Face*, edited and introduced by James Conant, Cambridge (Mass.) and London, Harward University Press.
 ----- (1992a): *Renewing Philosophy*, Cambridge (Mass.) and London, Harvard Univerty Press. Trad. esp.: *Cómo renovar la filosofía*, Madrid, Cátedra, 1994.
 ----- (1992b): “Replies”, *Philosophical Topics*, 20, 1 (1992), 347-408.
 ----- (1992c): *Il pragmatismo: una questione aperta*, Roma-Bari, Editori Laterza. Trad. esp.: *El pragmatismo. Un debate abierto*, Barcelona, Gedisa, 1999.
 ----- (1992d): “Atando cabos”, *Diánoia*, 1-15. V. In. en Putnam (1994b), 279-94.
 ----- (1993): “The Question of Realism”, en Putnam (1994b), 295-312.
 ----- (1994a): “Afterthoughts on ‘Models and Reality’” / “Nuevas ideas sobre ‘Los modelos y la realidad’”, *Diálogos*, 63 (1994), pp. 40-45.
 ----- (1994b): *Words & Life*, edited by James Conant, Cambridge (Mass.) and London, Harvard University Press.
 ----- (1994c): “Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind”, *The Journal of Philosophy*, XCI (1994), 445-465.

⁴ Los trabajos que han sido primero expuestos como conferencias, se han citado por el año en que se impartieron éstas, para que se viera mejor la evolución del pensamiento del autor.

----- (1997): "La importancia del conocimiento no científico", *Limbo/Teorema* XVI/2 (1997), 1-17.

José Antonio Guerrero del Amo
Universidad Complutense de Madrid
e-mail: jaguero@psi.ucm.es





KUHN Y EL PROBLEMA DE LA ELECCIÓN RACIONAL DE TEORÍA

GERMÁN GUERRERO PINO
Universidad del Valle

ABSTRACT: The communications shows and analyses what I consider is the Kuhn's principal thesis about rationality of theory choice: the rejection a logic of the evaluation theories. For to get preceding this thesis situate and delimit into context more extensive of the Kuhn's reflections: the notions of normal science and extraordinary science, and the relation between incommensurability and theory choice. Finally, I argue against Kuhn's idea what the science is pursuit of puzzle solving and I propose an ordering of virtues of a good theory in agreement with van Fraassen's view point.

KEYWORDS: Criteria of choice, evaluation theories, paradigm, puzzle-solving, incommensurability.

El problema de la elección racional de teoría, en lo que tiene de específico, sin lugar a dudas se relaciona con problemas más generales y con otros más específicos presentes en la reflexión filosofía de la actividad científica; problemas a los que Kuhn presentó soluciones alternativas novedosas respecto a las dominantes en lo que se suele llamar la *concepción heredada*. Esta situación se complica aún más si se tiene en cuenta los cambios que sufrió el punto de vista de Kuhn respecto a la elección de teoría. De modo que como estrategia de exposición es conveniente asumir una presentación sistemática más que histórica. Será sistemática ya que se concentra en las principales tesis presentadas y desarrolladas por Kuhn que tienen que ver directamente con la elección racional de teorías, y dejará de ser histórica en la medida en que presta poca atención al contexto en el cual se desarrollaron sus ideas y a las variaciones en su punto de vista.

Los dos primeros numerales se concentran especialmente en ubicar y delimitar las principales ideas sobre la elección de teorías dentro del contexto más amplio de las reflexiones de Kuhn: el modelo de desarrollo científico dominando por las ciencias normal y extraordinaria, y la relación entre incommensurabilidad y la evaluación de teorías. Este contexto permite

exponer y analizar, en el tercer numeral, lo que considero es la tesis metodológica principal de Kuhn: el rechazo de una lógica de evaluación de teorías. El último numeral problematiza la idea, defendida por Kuhn, de que la ciencia es una actividad de resolución de enigmas, proponiendo una jerarquía en el conjunto de criterios de elección a la luz de los planteamientos de van Fraassen.

Ciencias normal y extraordinaria

Para comenzar se tiene la tesis general del modelo de desarrollo científico que sustenta de manera importante las diferentes ideas presentadas por Kuhn. De acuerdo con Kuhn, la compleja actividad científica es susceptible de ser mejor entendida si se asume que el desarrollo científico comporta básicamente dos etapas (estadios o momentos): la ciencia normal y la ciencia o investigación extraordinaria. Kuhn en *La estructura de las revoluciones científicas* (1962) caracterizó estas dos etapas en función de términos como “paradigma” y “revolución”: “la ciencia normal como la actividad dominada por un paradigma” y “la ciencia extraordinaria como un periodo revolucionario en el que finalmente el paradigma dominante es sustituido”. Afinando más el análisis terminológico, puede notarse que el concepto de revolución científica depende del de paradigma científico, puesto que el primero precisamente se entiende a partir del segundo: como sustitución de un paradigma. Por tanto, la clarificación de las dos etapas predominantes en el desarrollo de una disciplina científica depende en buena medida de lo que se entienda por paradigma y, a su vez, por sustitución de un paradigma.

Pero la noción de “paradigma” que Kuhn presentó en su famoso libro es vaga y ambigua, dada la multiplicidad conceptual que otorgó a éste término. Kuhn trató de remediar esta dificultad en el *Postscript* de 1969 a la segunda edición de *La estructura de las revoluciones científicas* introduciendo el concepto de “matriz disciplinar” que recoge, a su parecer, las cuatro componentes esenciales de un paradigma: generalizaciones simbólicas, ejemplares o paradigmas propiamente dichos, valores metodológicos y modelos heurísticos u ontológicos; pero lo cierto es que este último término no arraigó y aún se sigue empleando el término “paradigma” en los círculos no especializados con la misma vaguedad característica. El mismo Kuhn a partir de los años setenta deja de emplear estos dos términos.

De modo que si nos atenemos más al espíritu que a la letra de los planteamientos kuhnianos, considero que la intuición principal que se encuentra detrás de la noción de paradigma es una nueva y compleja concepción de teoría empírica¹ en contraposición con la concepción de las teorías como un conjunto de enunciados, dominante en la primera mitad del

¹ Esta tesis está ampliamente sustentada en Moulines, 1982, pp. 61-77.

siglo. Esto por lo menos es lo que, por una parte, sugiere Kuhn en el *Postscript*² y, por otra, se puede concluir del hecho de que Kuhn en sus escritos posteriores hable más bien en términos de teorías empíricas que de paradigmas. De modo que para los propósitos presentes entonces sólo basta destacar que una teoría empírica es más compleja que un conjunto de axiomas y que mínimamente los dos elementos principales estrechamente vinculados que la componen son las leyes fundamentales (que contienen las generalizaciones simbólicas) y los ejemplares paradigmáticos o aplicaciones empíricas³.

Con este concepto de teoría empírica, sumado a la importante tesis de que las teorías evolucionan en el tiempo, podemos aproximar una mejor caracterización de “ciencia normal” y de “investigación extraordinaria”. Que las teorías empíricas evolucionen en el tiempo quiere decir que estas no se presentan completamente acabadas cuando se proponen y aceptan, sino que requieren de tiempo para su ajuste, desarrollo y ampliación –periodo de “ciencia normal”-. Además, y en pocas palabras, las actividades de ajuste, desarrollo y ampliación de una teoría inicialmente aceptada constituyen lo que Kuhn califica como *resolución de enigmas o rompecabezas* (“*puzzle-solving*”)⁴. De modo que la comunidad de científicos durante los periodos de ciencia normal se dedican de manera exclusiva a la resolución de enigmas, a resolver problemas teóricos, conceptuales, experimentales y metodológicos, sugeridos directa o indirectamente por la teoría. Si bien este tipo de trabajo normal produce modificaciones en la teoría en la que se está trabajando, estas modificaciones no son de carácter sustancial dado que las leyes fundamentales y aplicaciones paradigmáticas se mantienen; de ahí que se pueda decir que la comunidad científica trabaja en la misma teoría.

Las revoluciones científicas, por tanto, se presentan cuando una teoría es sustituida completamente por otra y vienen precedidas por un periodo de *investigación extraordinaria* en donde se configura de forma general una teoría alternativa a la dominante. Además, de acuerdo con Kuhn, la investigación extraordinaria es una actividad dominada por la búsqueda de solución a las *anomalías* empíricas de la teoría dominante, a experiencias o fenómenos que no encajan en la teoría; es decir, una anomalía es una contrastación desfavorable o negativa de la teoría que adquiere el carácter de anomalía por ser persistente dentro de la comunidad científica.

Se dice que se está frente a una nueva teoría en la medida en que las leyes fundamentales y, desde luego, las aplicaciones paradigmáticas de la

² Kuhn, 1983, p. 279.

³ Kuhn, 1983, p. 288, destaca esta misma idea dentro del contexto de aprendizaje de una teoría más que en el estructural de las teorías. Ver también Kuhn, 1977, pp. 321-331.

⁴ Kuhn otorgó mucha importancia a esta idea dentro de su concepción metodológica de la ciencia, ya que siempre estuvo dispuesto a defender que el objetivo de la ciencia es la “resolución de enigmas”. Sobre esta idea se vuelve al final del escrito.

teoría antigua han variado de manera considerable. Es por esto que las modificaciones en la teoría en caso de una revolución científica son modificaciones sustanciales y no accidentales como las presentes en los periodos de ciencia normal. De acuerdo con Kuhn, es en este contexto de teorías sucesivas que pretenden dar cuenta más o menos del mismo tipo de fenómenos donde se presentan los problemas de inconmensurabilidad y las dificultades de la elección de teoría.

Inconmensurabilidad y elección de teoría

Con las aclaraciones anteriores el problema de la elección racional de teoría puede formularse así: ¿cuál es el proceso que siguen los científicos para elegir entre teorías sucesivas e inconmensurables? Para comenzar, considero importante delimitar la relación existente entre inconmensurabilidad y elección de teoría en la concepción kuhniana, dado que es común exagerar dicha relación –actitud propiciada en parte por Kuhn, como a continuación veremos– trasladando las principales dificultades de la inconmensurabilidad al campo metodológico de la elección de teoría. Los profundos problemas filosóficos que encierra la relación de inconmensurabilidad entre teorías, sobre los cuales el mismo Kuhn llamó insistentemente la atención y que consideró constituyan el aspecto principal y novedoso de su propuesta filosófica, no radican propiamente en lo que concierne a la elección de teoría sino más bien se encuentran en el dominio epistemológico y semántico de la filosofía de la ciencia⁵, en temas como el realismo asociado directamente con la verdad de las teorías científicas y la variación ontosemántica en el cambio teórico, temas que actualmente se siguen debatiendo.

Kuhn en *La estructura de las revoluciones científicas* no presenta una noción de inconmensurabilidad suficientemente clara aunque sí al menos intenta delimitar, sin mayor precisión, tres clases de inconmensurabilidad: semántica, observacional y metodológica⁶. La primera clase de inconmensurabilidad, de acuerdo con las últimas versiones de Kuhn y que resultará ser la más pertinente en los análisis diacrónicos de la ciencia, consiste en “que no hay ningún lenguaje, neutral o de cualquier otro tipo, al que ambas teorías, concebidas como conjuntos de enunciados, puedan

⁵ Este punto de vista también lo defiende Ernan McMullin, 1995, pp.55 y 56.

⁶ Ver Kuhn, 1983, pp. 230-234. Kitcher, 1982, pp. 690, destaca también estas tres clases de inconmensurabilidad aunque a la primera la llama “conceptual”. Entre tanto, Newton-Smith, 1987, pp. 126-132, 166 y 167, no habla de inconmensurabilidad observacional, pero sí de las otras dos, dividiendo la inconmensurabilidad metodológica en dos: variación de valor y variación radical de patrón. Esta última es la que se quiere destacar con la metodológica.. Como veremos en el siguiente numeral, lo que aquí se ha calificado como inconmensurabilidad observacional es más adecuado caracterizarlo como la dependencia teórica de la observación, que constituye una tesis epistemológica importancia dentro de la concepción de Kuhn y de la cual depende en buena medida su tesis metodológica de la imposibilidad de una lógica de la evaluación de teorías.

traducirse sin resto o pérdida. Ni en su forma metafórica ni en su forma literal inconmensurabilidad implica incomparabilidad”⁷. Es a este tipo de variación semántica entre teorías sucesivas a la que se referirá Kuhn, a partir de los años ochenta y de manera exclusiva, como una relación de inconmensurabilidad. La inconmensurabilidad observacional, que será objeto de una mayor precisión en el numeral siguiente, tiene que ver con la idea de la dependencia teórica de la observación, tesis epistemológica central en el giro historicista de la filosofía de la ciencia llevado a cabo en la década de los sesenta por Kuhn, Feyerabend y otros. La tercera clase, la que más interesa en el presente contexto, tiene que ver con los criterios o valores epistemológicos involucrados en la elección de teorías.

Lo que se defiende aquí, entonces, es que Kuhn al plantear este tipo de inconmensurabilidad metodológica está apoyando una tesis radical en cuanto a la evaluación de teorías, que bien podría calificarse como “relativismo metodológico”, puesto que sus afirmaciones y argumentos más determinantes apuntan hacia aquella tesis en donde la elección entre paradigmas no puede estar determinada por factores internos -precisión, fecundidad, coherencia, amplitud, simplicidad, etc.- sino por los estrictamente externos o si se quiere, de manera más relajada, que la función de los primeros es secundaria en la evaluación de teorías⁸. Para complementar la idea de que los criterios internos se encuentran en un segundo plano respecto a los factores externos, sólo basta ver cómo Kuhn hace depender los valores del paradigma que se defiende; es decir, los criterios de elección varían categóricamente de un paradigma a otro o, lo que es lo mismo, de una teoría a otra. Esto es claro puesto que su mismo concepto de paradigma, posteriormente precisado en la *Postscript* a través del de “matriz disciplinar”, involucra los valores como una de sus componentes, de modo que al variar los valores de un paradigma a otro y dada la inconmensurabilidad entre estos, los valores mismos estarían sometidos a una suerte de inconmensurabilidad en donde la disputa entre paradigmas tiene que ser resuelta por factores externos⁹.

Posteriormente Kuhn en el *Postscript* y, de manera más consistente, en *Objetividad, juicios de valor y elección de teoría* (1973), presenta una tesis menos radical puesto que la función de los criterios internos de elección ya no es secundaria, como inicialmente lo pensaba, sino que de alguna forma influyen en, más no determinan, la elección de teoría. Además, la tesis es menos radical porque se reduce la dependencia de dichos criterios respecto a un paradigma, y en la misma proporción las dificultades que la

⁷ Kuhn, 1989, p. 99.

⁸ Esto es particularmente notable si se compara Kuhn, 1983, p. 175, con Kuhn, 1983, pp. 237-244.

⁹ Ver Kuhn, 1983, p. 152.

incommensurabilidad de teorías inyectaba a la evaluación de teorías¹⁰. En muchos lugares Kuhn expresa su distanciamiento de su tesis radical inicial; por ejemplo, en la *Postscript*, al referirse a los valores como a una de las componentes de la matriz disciplinar plantea que “habitualmente se les comparte entre diferentes comunidades...y hacen mucho para dar un sentido de comunidad a los científicos naturalistas en conjunto... Considero ahora como una flaqueza de mi texto original el haber prestado poca atención a valores tales como la coherencia interna y externa al considerar las causas de crisis y factores de elección de teorías”¹¹.

Una vez precisado el vínculo entre la tesis de la incommensurabilidad y la evaluación de teorías, podemos enunciar la tesis que finalmente Kuhn defiende en cuanto a la elección de teorías en los siguientes términos: el conjunto de criterios o valores de evaluación de teorías que guían a (pero no dictan a) los científicos es relativamente estable en el tiempo y de una comunidad a otra, aunque la manera de entender y aplicar cada uno de ellos, como la importancia relativa entre los mismos, puede cambiar de manera significativa de un científico a otro, aún compartiendo la misma teoría¹². Por cuestión de espacio no es posible analizar cada una de las partes que constituyen esta tesis, por tanto sólo nos concentraremos en la que considero es la más determinante - el conjunto de valores epistemológicos de evaluación de teorías guían pero no dictan la elección- de modo que las otras aparecerán en un segundo plano.

Evaluación de teorías

La anterior subtesis, que igualmente se puede considerar como una tesis más restringida, ha sido una constante en líneas generales en las diversas reflexiones de Kuhn sobre el tema y, en pocas palabras, la tesis lo que propone es que no hay una lógica subyacente en los procesos de evaluación o elección de teorías¹³. Esta tesis es una crítica frontal a aquellos planteamientos del empirismo lógico y del racionalismo crítico de Popper que apoyan, explícita o implícitamente, la idea de una lógica en la evaluación de las teorías en el sentido de que una teoría se abandona cuando de ella se deduce una contradicción interna o se enfrenta con una experiencia refutadora.

En los casos históricos analizados por Kuhn se evidencia que los científicos están dispuestos a mantener una teoría incluso en presencia de confirmaciones desfavorables y de contradicciones internas, siempre y cuando

¹⁰ La tesis que aquí se está presentando la comparte igualmente Newton-Smith, pues dice: “el propio Kuhn ha expresado su acuerdo en los últimos tiempos, a saber, la del consenso con independencia de los paradigmas en lo relativo a las características que hacen de una teoría una buena teoría”. Newton-Smith, 1987, p. 129.

¹¹ Kuhn, 1983, p. 283.

¹² Kuhn, 1977, p. 360.

¹³ Ver especialmente Kuhn, 1975, p. 521, para una formulación precisa.

esta situación no se haga crítica y generalizada. Esta es la actitud característica de los científicos en los periodos de ciencia normal dada la confianza que tienen en la teoría que están implementando, pero esta confianza se reduce a medida que se agudizan las contradicciones, internas o con la experiencia, lo que da lugar a la búsqueda de nuevas alternativas teóricas, a las revoluciones científicas.

Además, la tesis anterior de Kuhn se encuentra soportada, en parte, por dos tesis epistemológicas que poseen un carácter más general: la dependencia teórica de la observación y el holismo metodológico. De acuerdo con la primera, y en pocas palabras, no existe un lenguaje de observación totalmente neutral a dos teorías sucesivas. Es decir, los datos de la experiencia no son neutrales sino que su interpretación depende de hipótesis teóricas implícitas que conceptualizan la experiencia. De modo que ante una confirmación desfavorable cabe la posibilidad de revisar este tipo de hipótesis antes que revisar la teoría misma.

Por su parte, la tesis de Duhem-Quine, una forma de holismo metodológico que parte de concebir las teorías como un conjunto de hipótesis, plantea que ante una contradicción con la experiencia lo que resulta refutado en sentido estricto no es una hipótesis particular de la teoría sino la teoría en su totalidad. Por tanto, el proceso de revisión de la teoría puede darse de múltiples formas modificando una o varias de las hipótesis componentes. Pero el holismo metodológico defendido por Kuhn adquiere una forma diferente a la anterior dada la “concepción compleja” de teoría implícita en sus planteamientos -tal como se ha tratado de mostrar más arriba-. De acuerdo con la versión del holismo metodológico de Kuhn, y partiendo de aceptar una experiencia refutadora, la lógica no obliga a abandonar la teoría puesto que la revisión del sistema teórico puede hacerse por dos caminos que llevan a resultados distintos: cambiando una o varias hipótesis específicas sin modificar sus leyes fundamentales, de modo que en sentido estricto nos encontramos ante la misma teoría; o cambiando radicalmente de teoría, modificando una o varias de las leyes fundamentales.

En síntesis, de acuerdo con lo dicho hasta aquí, el proceso de elección puede ser dividido en las siguientes cuatro etapas: ante la presencia de anomalías empíricas un número reducido de científicos trabajan en la elaboración de una alternativa teórica, asumiendo como foco principal de investigación la solución de las anomalías; una vez encontrada una teoría alterna, mínimamente desarrollada y explotada, sus proponentes destacan un mínimo de ventajas en relación con la teoría dominante; la teoría alterna se somete a la crítica de quienes vienen trabajando en la teoría dominante; finalmente se acumulan tal cantidad de argumentos a favor de la nueva teoría que, aun cambiando las razones de aceptación de un miembro a otro, casi todo los miembros hacen la misma elección. Además, los criterios epistemológicos -coherencia interna y externa, precisión, amplitud, etc.- no operan como reglas

de aplicación mecánica, como algoritmos que *imponen y determinan* de manera rígida la elección de una teoría, sino que por el contrario dichos criterios verdaderamente se emplean como valores que *influyen* en la decisión. De manera aún más general, se rechaza que la racionalidad científica esté regida de manera exclusiva por cánones lógicos, no hay duda de que la racionalidad también depende de factores pragmáticos presentes en algunos de los valores epistemológicos.

Jerarquía en los criterios de elección

El punto que deseo discutir ahora tiene que ver con la posibilidad de establecer alguna clase de orden entre estos criterios principales y, de ser posible, determinar cuál sería la más indicada. Considero que Kuhn rechazó tal posibilidad y más bien defendió cierto tipo de “democracia” en la función de los criterios de elección, puesto que estos están en igualdad de condiciones en lo que respecta a los diferentes grados de influencia que pueden tener en la evaluación de teorías. Esto es así puesto que Kuhn propone que, en primer lugar, los criterios de elección forman un campo semántico¹⁴ que permite dar sentido al término moderno “ciencia”, pero de un modo tal que este campo semántico disciplinar no puede entenderse como un conjunto de criterios que estipula las condiciones necesarias y suficientes para que una disciplina sea calificada como científica. Los criterios simplemente resultan útiles para calificar a algunas disciplinas como científicas, a otras como no científicas y a otras como estando en el limbo. En segundo lugar -y en esto Kuhn sigue a Hempel¹⁵- dado que lo común ha sido expresar los objetivos de la actividad científica en términos de los medios que permiten alcanzarlos -las supuestas virtudes de una buena teoría-, si se aplica un principio metodológico de simplicidad no es necesario diferenciar entre los primeros y los segundos sino que es más adecuado asumir los criterios de elección como los fines mismos de la actividad científica. En tercer lugar -Kuhn simplifica más la exposición-, dichas virtudes de una buena teoría son “las características que definen la resolución de un enigma”¹⁶ y, por tanto, “la actividad de resolución de enigmas es más que suficiente para hacer de ella un fin en sí mismo de la ciencia”¹⁷.

¹⁴ Kuhn, 1989, p. 148.

¹⁵ Kuhn, 1989.

¹⁶ Kuhn, 1995, p. 338.

¹⁷ Ibid. Además Kuhn, 1994, pp. 34 y 35, plantea que “los juicios de esta clase son necesariamente comparativos: cuál de dos cuerpos de conocimiento -el original o el propuesto como alternativo- es *mejor* para hacer lo que hacen los científicos, cualquier cosa que ello sea. Y ello es así tanto si lo que hacen los científicos es resolver enigmas (mi punto de vista), mejorar la adecuación empírica (el de Bas van Fraassen), o incrementar la dominación de la élite gobernante”.

La idea, entonces, que se quiere presentar aquí, inspirada principalmente en los planteamientos de van Fraassen, es que virtudes de una buena teoría como la coherencia interna -ausencia de contradicciones- y el poder predictivo – mayor precisión en las predicciones- desempeñan funciones más determinantes en los procesos de elección respecto al resto de virtudes y que, por tanto, la máxima de Kuhn –la ciencia ante todo como una actividad de resolución de enigmas- deja fuera u oculta en buena forma dicha distinción. Además, puesto que el desarrollo detallado de esta idea desborda los límites del presente escrito simplemente la presentaré en sus líneas generales queriendo destacar especialmente que estas dos virtudes epistemológicas tienen la función de ser criterios fundamentales de aceptabilidad mínima. De modo que, dado que valores como la verdad y la mejor explicación se encuentran entre los más aplaudidos y constituyen el núcleo de controversias importantes y profundas, las objeciones presentadas a aquellos puntos de vista que defienden ya sea que el objetivo de la ciencia es “la búsqueda de la verdad” o “la búsqueda de mejores explicaciones” remito a los escritos de Kuhn para el primer caso y a los de van Fraassen para ambos casos, pues considero que por lo menos las objeciones de ambos al realismo convergen en puntos importantes.

Hechas estas salvedades lo primero a subrayar es que en lo que concierne a las virtudes de una buena teoría se encuentra, de acuerdo con van Fraassen, una división tripartita de propiedades y relaciones¹⁸. Propiedades de una teoría como la axiomatibilidad y la consistencia –ausencia de contradicciones- corresponden a relaciones puramente internas o lógicas; en tanto que propiedades como la verdad y el poder predictivo –adecuación empírica- son relaciones semánticas que conciernen a la relación entre la teoría con el mundo; y, finalmente, hay propiedades teórico – pragmáticas como simplicidad, fecundidad, amplitud, etc., que tienen que ver con relaciones entre la teoría y sus usuarios, dependiendo así radicalmente del contexto.

Ahora bien, esta diversidad de funciones contiene en sí misma un orden el cual permite dar mayor prioridad a unos criterios que a otros. Es en este sentido en el que van Fraassen plantea que la consistencia interna y la adecuación empírica son criterios fundamentales de aceptabilidad mínima, pues independientemente de las otras virtudes no se puede aceptar una teoría que sea empíricamente adecuada e inconsistente; e igualmente no se puede aceptar una teoría que siendo consistente internamente tenga contradicciones con experiencias previamente establecidas. Mientras que propiedades como simplicidad, fecundidad y amplitud no tienen esta peculiaridad; es decir, por lo menos en teoría, no tiene sentido valorar a una de ellas independientemente de la consistencia interna y el poder predictivo de la teoría.

¹⁸ van Fraassen, 1996, p. 117 y 118.

Pero este tipo de jerarquía en los méritos que asignamos a una buena teoría es compatible con la idea de que los mismos tomados en conjunto forman un sistema de valores epistemológicos que funciona como un todo en la evaluación de teorías, ya que “en el análisis de la evaluación de las teorías científicas – plantea van Fraassen- sería un error soslayar las maneras en las cuales esa evaluación está teñida de factores contextuales. El científico lleva estos factores a la situación desde su propia situación personal, social y cultural. Es un error creer que los términos en los que se aprecia una teoría científica son puramente higiénicos, y no tienen nada que ver con ningún otro tipo de valoración, o con las personas y las circunstancias involucradas”¹⁹

Finalmente, haber puesto de relieve que las propiedades de una teoría contienen diferentes tipos de relación es ya de por sí dar un paso adelante en el análisis en relación con el estado en que lo dejó Kuhn. Además, considero que algunos pasajes de la obra de Kuhn sugieren que tanto la consistencia interna como la adecuación empírica deben ser considerados criterios fundamentales en el sentido aquí expuesto, aunque la idea que explícitamente defiende, como ya se ha dicho, no busca hacer tales distinciones. Por ejemplo, Kuhn finaliza su artículo *Racionalidad y elección de teorías*, de manera un tanto enigmática y definitiva, así: “un requisito para cualquiera de los dos (‘racionalidad’ y ‘justificación’) es ajustarse a las limitaciones impuestas por la lógica. Otro requisito es ajustarse a las limitaciones impuestas por la experiencia en ausencia de buenas razones para lo contrario. Ambos requisitos ponen de manifiesto parte de lo que implica ser racional”²⁰.

BIBLIOGRAFÍA

- Kitcher, Philip: “Implications of Incommensurability”, en Asquith, P.D., Nickles, T., *PSA 1982, Vol. 2*, Michigan, East Lansing, pp. 689-703.
- Kuhn, Thomas S.: *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press, 1996. (Vers. Cast. *La Estructura de las Revoluciones Científicas*, México, F. C. E., 1983).
- “Notas sobre Lakatos”, en *La Crítica y el Desarrollo del Conocimiento*, Barcelona, Grijalbo, 1975.
 - “Algo más sobre los paradigmas”, en *La tensión esencial*, México, F. C. E., 1977, pp. 317-343.
 - “Objetividad, juicios de valor y elección de teorías”, en *La tensión esencial*, México, F. C. E., 1977, pp. 344-364.
 - “Rationality and Theory Choice”, *The Journal of Philosophy*, v.80, n.10, 1983, pp.563-570. (Vers. Cast. *¿Qué son las Revoluciones Científicas? y otros Ensayos*, Barcelona, Paidós, 1989, pp. 137-151).
 - “Commensurability, Comparability, Communicability”, en P.D. Asquith and T. Nickles (comp.), *PSA 1982, Volume 2*, East Lansing, Michigan, Philosophy of Science Association, pp. 669-688. (Vers. Cast. *¿Qué son las Revoluciones Científicas? y otros Ensayos*, Barcelona, Paidós, 1989, pp. 95-135).

¹⁹ Ibid, p. 114.

²⁰ Kuhn, 1989, p. 151. Ver también pp. 138 y 139 del mismo escrito y el *Postcrip*, p. 283.

- "The Road Since Structure", en A. Fine, M. Forbes and L. Wesseles (comp.), *PSA 1990, Volume 2*, East Lansing, Philosophy of Science Association, 3-13. (Vers. Cast. El Camino desde "La Estructura", *Arbor CXLVIII*, 583, Madrid, Agosto 1994, pp. 27-46)
 - "Afterwords", en P. Horwich (comp.), *World Changes: Thomas Kuhn and the Nature of Science*, Cambridge, MIT Press, 1995, pp. 311-341.
- McMullin, Ernan: "Rationality and Paradigm Change in Science", en P. Horwich (comp), *World Changes: Thomas Kuhn and the Nature of Science*, Cambridge, MIT Press, 1995, pp. 55-78.
- Moulines, C. Ulises: *Exploraciones Metacientíficas*, Madrid, Alianza, 1982.
- Newton-Smith, W.H.: *La Racionalidad de la Ciencia*, Barcelona, Paidós, 1987.
- Van Fraassen, Bas: *La Imagen Científica*, México, Paidós y Universidad Nacional Autónoma de México, 1996.

Germán Guerrero Pino
Universidad del Valle
e-mail: gegerrerr@retemail.es





VERDAD Y ÉXITO EMPÍRICO¹

VALERIANO IRANZO

ABSTRACT: The aim of this paper is to analyse the explanationist defense of scientific realism. According to it, truth is the best explanation of science's empirical success. Instrumentalist and pragmatist accounts explain why we arrive to empirically successful theories whilst scientific realism explains why a theory is empirically successful. Since there are two different *explananda* involved, I see no rivalry between both approaches. A sociological account of science (Strong Programme, EPOR), in contrast, is not compatible with the explanationist defense of scientific realism insofar as it rejects any extra-subjective value for the notion of empirical evidence. I point out some aspects of scientific practice which collide with the sociological approach. Yet I conclude that, even though truth is the best explanation of why a theory is empirically successful, it is not a very good one by now because of successful theories discarded long time ago.

KEYWORDS: Scientific realism, truth, empirical success, explanation, explanation (inference to the best).

La detección de una radiación uniforme, de baja intensidad, proveniente de todas las direcciones del espacio, fue crucial para confirmar la hipótesis del Big-Bang. La ausencia de explicaciones alternativas de este fenómeno y la confirmación de ulteriores predicciones nos llevan a creer que la hipótesis es verdadera, es decir, que realmente se produjo una rapidísima expansión del Universo hace unos quince mil millones de años. Remedando a Putnam, la atribución de verdad es la única posibilidad para no convertir el éxito de la hipótesis en algo milagroso.

El argumento "explicacionista" a favor del realismo científico sostiene, pues, que la verdad de la teorías es la mejor explicación –si no la única, según algunos autores- de su éxito. Aunque ésta es una cruda formulación del

¹ La investigación conducente a esta ponencia ha sido financiada por el Ministerio de Educación y Cultura mediante una beca del Programa de Perfeccionamiento de Doctores en el extranjero. Agradezco también a la Dirección General Científica y Técnica (DGICYT) su subvención como parte del proyecto PB 96-0764.

argumento -Richard Boyd o Jarrett Leplin, por mencionar sólo dos ejemplos, han ofrecido versiones más sofisticadas²-, el denominador común a todas ellas es la conexión entre éxito y verdad: la verdad de la hipótesis teórica se infiere a partir de los éxitos que ésta cosecha. En esta ponencia pretendo discutir los supuestos de este argumento. Considero que es una buena estrategia de cara a valorar las posibilidades de bloquearlo desde el frente antirrealista. No obstante, primero haré unos breves comentarios sobre las nociones de verdad y éxito.

Que la teoría es verdadera significa para el realista que la parcela de realidad a la que la teoría se aplica con éxito es –aproximadamente al menos– como la teoría describe. Seguro que algunos –o muchos– de ustedes tendrán sus escrúpulos acerca de una noción de verdad tal. Pero no se trata ahora de comprometerse con ella, sino sólo de apuntar la noción de verdad que el realista ha de inferir para poder rentabilizar el argumento.

Por otro lado, y salvo excepciones, no hay una caracterización precisa de lo que los autores realistas entienden por éxito. El acento suele recaer bien en el éxito tecnológico-manipulativo –la “fiabilidad instrumental” de la metodología científica, según Boyd–, bien en el éxito predictivo –las predicciones “novedosas”, de Leplin. Una discusión del argumento realista como la que nos ocupa nos obliga a pasar por alto las particularidades introducidas por autores concretos. Por eso creo que bastará con entender éxito como el cumplimiento de las expectativas predictivo-instrumentales generadas por la teoría. Puesto que se trata del éxito logrado por la teoría en nuestro intercambio con la experiencia, hablaré simplemente de *éxito empírico*. Desde luego, una teoría empíricamente exitosa no es, simplemente, una teoría capaz de encajar la evidencia disponible en un momento determinado. El cuerpo evidencial relevante para una teoría científica constituye un dominio en ampliación, ampliación no sólo cuantitativa, ya que muy a menudo nuevos instrumentos y nuevas técnicas posibilitan acceder a fuentes de información previamente inaccesibles. Por eso el éxito empírico, tal como lo entenderemos aquí, exige la capacidad *sostenida* de dar cuenta de la evidencia³.

I

El argumento realista asume los siguientes supuestos:

(S₁) El éxito empírico de las teorías científicas requiere una explicación.

² Jarret Leplin, *A Novel Defense of Scientific Realism*, Oxford, Blackwell, 1997; Richard Boyd, “Realism, Approximate Truth and Method”, en Papineau, David, ed., *The Philosophy of Science* (Oxford, Oxford U.P., 1996), pp. 215-55.

³ La idea de que el éxito predictivo puntual no es, por sí mismo, un indicador fiable es un lugar común para muchos realistas. V., por ejemplo, “Rationality and Paradigm Change in Science”, Ernan McMullin, en *World Changes: Thomas Kuhn and the Nature of Science*, Horwich, P., ed., Cambridge, Mass., MIT Press, 1993.

(S₂) La verdad explica el éxito empírico.

(S₃) La mejor explicación del éxito empírico es que las teorías empíricamente exitosas son (aproximadamente) verdaderas.

(S₄) La mejor explicación de un hecho es verdadera (o aproximadamente verdadera, o altamente probable).

Comentaré S₁, S₂ y S₃, por este orden. Dejaré a un lado S₄, pues su discusión plantea tópicos, como la legitimidad de la *inferencia abductiva* o el valor epistémico de la bondad explicativa, que nos llevarían muy lejos.

(S₁) *El éxito empírico de las teorías científicas requiere una explicación*

El antirrealista puede negar esta afirmación arguyendo que el afán por explicar el éxito empírico es una necesidad que sólo siente el realista. Pero aunque esto sea así, el realista puede defender S₁ eliminando el componente psicológico. En principio, todo hecho que realmente tiene o ha tenido lugar *requiere* una explicación. El término ‘requiere’ no alude, pues, a la compulsión psicológica que provoca en nosotros un hecho particular, compulsión que nos lleva a la búsqueda de una explicación para calmar el desasosiego cognitivo, sino a una de nuestras creencias básicas respecto a la realidad, a saber, que la ocurrencia de un hecho depende de algún otro hecho. En este sentido, todos los hechos ocurridos, tanto el más asombroso como el más anodino, requieren, en principio, una explicación. Por consiguiente, en tanto el éxito empírico sea un hecho contrastado, requiere una explicación.

Una posible objeción al realista sería su parcialidad a la hora de enjuiciar los resultados cosechados por la ciencia. Sabemos que muchos proyectos de investigación que en sus inicios fueron considerados prometedores posteriormente acabaron convirtiéndose en callejones sin salida y hoy sólo tienen interés para el historiador. El camino adecuado se ha discernido, en ocasiones, tras una serie de insatisfactorios tanteos, y algunos de los éxitos conseguidos han sido efectos imprevistos de la investigación emprendida. ¿No será que lo que el realista trata de explicar es una ilusión consecuencia de una visión parcial y triunfalista de la ciencia?

Esta objeción es peligrosa para el realista tanto si implica la negación directa del *explanandum* (“no hay éxito empírico”), como si convierte el éxito empírico en un fenómeno tan inusual que no podemos considerarlo sino una mera casualidad (“puesto que ha habido muchos intentos, algún acierto teórico se habrá producido”).

Pienso, no obstante, que ninguna de ambas conclusiones se sigue de admitir que ha habido sonados fracasos en la historia de la ciencia. No veo por qué admitir esto es incompatible con el reconocimiento de que al menos ciertas parcelas de la ciencia contemporánea son altamente exitosas en cuanto a su rendimiento empírico, esto es, en cuanto al cumplimiento de las expectativas predictivo-instrumentales generadas por ellas. Y este punto de apoyo –difícil

de negar salvo desde un escepticismo o un relativismo radicales- es todo lo que necesita el realista para poner en marcha su argumentación. No hace falta que el número de éxitos sea abrumadoramente superior a la de fracasos; basta con que haya un número significativo de éxitos, aun a costa de un número también significativo de fracasos. Que sea posible llegar a la conclusión realista desde este punto de partida es harina de otro costal.

Así pues, explicamos, prioritariamente, hechos. Y puesto que el éxito empírico es un hecho contrastado, requiere una explicación. Sin embargo, requerir una explicación no es lo mismo que tenerla. Con otras palabras, puede haber hechos que, en tanto hechos, requieran una explicación y que, sin embargo, no la tengan. La relación entre la masa del protón y la del electrón es $1/1836$. Este hecho podría ser explicado si se encontrara, por ejemplo, un elemento constituyente común a ambos. Hoy por hoy este hecho no tiene explicación, aunque quizá algún día la tenga. La cuestión entonces es si el hecho que nos ocupa, el éxito empírico de la ciencia, tiene o no tiene explicación. Veamos alguna de las explicaciones propuestas.

Las hipótesis científicas son formuladas y seleccionadas teniendo en mente el éxito empírico. No resulta entonces sorprendente que haya hipótesis empíricamente exitosas, ya que se trata precisamente de un objetivo explícito en la elaboración de teorías. De acuerdo con este razonamiento, las teorías científicas son empíricamente exitosas porque las que resultan ser manifiestamente inadecuadas en cuanto a su rendimiento empírico no superan la criba. Sólo las que puntúan alto en este aspecto llegan a gozar del beneplácito de la comunidad científica⁴.

Cabe también una versión naturalizada e historicista del argumento, más refinada tal vez. En la ciencia tanto el nivel metodológico como el axiológico son moldeados por la experiencia. Frente a otros fines de la ciencia que han sido descartados por ser irrealizables, el éxito empírico es un objetivo que ha perdurado hasta la actualidad. Por otro lado, durante varios siglos la metodología científica se ha ido modelando para conseguir dicho fin. Como resultado, entre las teorías aceptadas por los científicos hemos de encontrar bastantes que sean exitosas⁵.

Ambas explicaciones destacan, en último término, el hecho de que la elaboración de teorías empíricamente exitosas es un fin de la actividad científica. Suponiendo que ésta es razonablemente efectiva en la consecución de dicho fin, la obtención de teorías empíricamente exitosas es así un resultado natural. Esta argumentación basada en la criba ejercida por la metodología resulta inicialmente plausible. Sin embargo, si el éxito empírico queda explicado

⁴ Bas van Fraassen, *The Scientific Image* (Oxford: Clarendon Press, 1980), pp. 39-40.

⁵ Larry Laudan, "Explaining the Success of Science: Beyond Epistemic Realism and Relativism", en Cushing, J., Delaney, C.F. y Gutting, G., eds., *Science and Reality*, Notre Dame, Ind., University of Notre Dame Press, 1984.

con ella, el realismo científico es, cuanto menos, redundante. ¿En qué sentido, pues, es deficitaria la explicación en términos de la “criba metodológica”? ¿Por qué es superior la explicación realista?

(S₂) *La verdad explica el éxito empírico*

Distingamos dos cuestiones respecto al éxito empírico de la ciencia. No es lo mismo preguntar cómo o por qué llegamos a elaborar teorías empíricamente exitosas que preguntar por qué dichas teorías son exitosas. La primera cuestión sí puede contestarse apelando a los mecanismos de control que operan en la ciencia, fundamentalmente la metodología y los criterios de selección entre teorías rivales, los cuales favorecen la obtención de teorías empíricamente exitosas. Mas con esto no hemos respondido a la segunda cuestión.

Supongamos dos hipótesis rivales, h_1 y h_2 , que encajan con la evidencia disponible y que ofrecen una explicación aceptable de ella. Supongamos también que h_1 anticipa evidencia ulterior mientras que h_2 no hace más que encajarla de un modo *ad-hoc*. Los controles anteriormente mencionados favorecen la adopción de h_1 . Apelar a ellos permite explicar por qué la comunidad científica acaba prefiriendo h_1 sobre h_2 , pero no explica por qué h_1 tiene éxito y h_2 no.

Aquí es donde el realista introduce su explicación. Para él el éxito empírico -el cumplimiento de las expectativas predictivo-instrumentales generadas por la hipótesis, tal como lo hemos definido- es consecuencia de su verdad. Al argumento realista subyace la intuición de que la efectividad en la anticipación de lo que va a ocurrir significa que la identificación y caracterización de las entidades involucradas en los fenómenos objeto de estudio es bastante aproximada. Precisamente eso es lo que explica por qué las cosas ocurren del modo en que ocurren, a saber, porque los mecanismos y procesos causales postulados por la teoría existen y están operando *grosso modo* del modo en que nuestra teoría dice que operan. Y a la inversa, el éxito empírico es un síntoma de que se ha diseccionado la naturaleza por sus junturas (*carving nature at its joints*). De este modo la verdad explica el éxito empírico, según el realista.

Ahora se entenderá mejor en qué sentido la verdad explica el éxito empírico y por qué la explicación realista no es ni redundante ni incompatible con la explicación que apela a la criba metodológica. Ambas pueden ser aceptadas a la vez porque el éxito empírico está involucrado en dos *explananda* diferentes. De ahí que no pueda decirse, sin más, que la explicación realista sea mejor. La explicación en términos de la criba metodológica es correcta, pero la explicación realista no es un aditamento superfluo pues, respecto a por qué

una teoría tiene éxito empírico, aquélla no nos dice nada⁶. Esta maniobra permite, en fin, precisar el aporte específico proporcionado por la explicación realista y reconocer al mismo tiempo lo que hay de plausible en la idea de la presión selectiva ejercida por la metodología.

S₃ La mejor explicación del éxito empírico es que las teorías empíricamente exitosas son (aproximadamente) verdaderas

Puesto que *S₃* afirma que la explicación realista es la mejor explicación del éxito empírico —en el aspecto mencionado—, una vez descartada la explicación de la presión selectiva, habrá que indagar si existe alguna otra alternativa con la que comparar.

Desde un planteamiento bien distinto al realismo se podría objetar que estamos entendiendo el éxito empírico como un factor en gran medida ajeno a la influencia del sujeto; sin embargo, ¿acaso no encontramos aquello que queremos encontrar? El científico realiza una selección, a menudo inconsciente, de la evidencia de modo que tiende a salvar lo que confirma su hipótesis. O bien maximiza los aciertos y minimiza los fracasos. El caso es que la valoración del rendimiento empírico de una hipótesis siempre es parcial. Por eso, el éxito empírico no resulta de la aquiescencia de la realidad, sino del sesgo introducido por el propio investigador en la confrontación con la experiencia. El resto es política científica. La credibilidad que disfrutará la hipótesis dependerá de la influencia que sus defensores sean capaces de ejercer sobre el resto de la comunidad científica, influencia basada en gran medida en factores extra-racionales.

Este enfoque de la actividad científica —en la línea de aproximaciones como el Programa fuerte (*Strong Programme*) o el Programa empírico del relativismo (*Empirical Programme of Relativism - EPOR*)⁷— sugiere invertir la relación entre éxito empírico y aceptación por parte de la comunidad científica. La teoría es empíricamente exitosa porque se acepta; pero no se acepta porque es exitosa. El éxito empírico no es más que un epifenómeno, un capítulo más en el proceso de negociación e imposición en que consiste la aceptación de una hipótesis científica. Si la evidencia que decide el éxito empírico de una teoría está completamente determinada por los intereses de individuos, grupos e instituciones, entonces, los triunfos de una teoría no son más que los triunfos de los intereses de sus partidarios. ¿Cómo considerar que el éxito empírico es un síntoma de que la teoría es aproximadamente verdadera, tal y como

⁶ En este sentido v. Leplin, *op. cit.*, cap. 1.

⁷ V. “Relativismo, racionalismo y sociología del conocimiento” de Barry Barnes y David Bloor, así como “Un programa empírico de relativismo en sociología del conocimiento científico”, de Harry Collins. Ambos artículos se encuentran en González García, M.I., López Cerezo, J.A. y Luján, J.L., eds., *Ciencia, Tecnología y Sociedad. Lecturas seleccionadas*. Barcelona, Ariel, 1997, pp. 27-48 y 49-66.

preconiza el realista?

En sentido estricto no estamos ante una explicación del éxito empírico. Lo que se nos está diciendo es que es equivocado pensar que el éxito empírico de la teoría depende de su apoyo evidencial, como supone el realista⁸. Aquí conviene destacar dos cosas. Primero, si la posición del defensor del Programa fuerte o del EPOR resulta plausible, la explicación realista no es una buena explicación, en tanto su concepción del *explanandum* no es correcta. Estamos, entonces, ante una alternativa al argumento realista, ya que resultaría incoherente aceptar éste conjuntamente con la redefinición “sociológica” del éxito empírico (mientras que sí es posible, como hemos visto antes, conjuntar la explicación realista y la explicación basada en la criba metodológica). Por eso, lo buena que sea la explicación realista depende de si la reinterpretación sociologista es viable o no. Segundo, la viabilidad de la reinterpretación sociologista dependerá a su vez de sus ventajas o limitaciones a la hora de dar cuenta de otros aspectos de la actividad científica distintos al éxito empírico, puesto que, en sentido estricto, éste no ha sido explicado. Sin ánimo de realizar un análisis exhaustivo, señalaré algunas consecuencias problemáticas de dicha reinterpretación.

Contra la tesis de que el éxito empírico no tiene valor epistémico, porque nunca hay un arbitraje neutral por parte de la experiencia, cabe defender la existencia de una evidencia preteórica que juega un papel importante en la resolución de las disputas científicas. Desde luego, esta evidencia preteórica no ha de entenderse como un sustrato experiencial puro, basta con que no presuponga las afirmaciones teóricas objeto de contrastación. El hecho de que, en ocasiones, un grupo investigador admita la evidencia que favorece a un grupo rival, e incluso pueda colaborar en obtenerla, debilitando así la propia hipótesis, es un indicio de la neutralidad que se confiere a la evidencia en la práctica científica⁹. Por otra parte, antes ya he aludido a los descubrimientos inesperados. No es difícil encontrar episodios de la ciencia en los que la evidencia obtenida no sirvió para confirmar la hipótesis que había motivado la investigación, aunque abrió otras posibilidades interesantes¹⁰.

⁸ Y no sólo el realista. Antirrealistas de talante tan dispar como van Fraassen o Laudan coinciden en este punto con el realista.

⁹ La primera predicción de la hipótesis del Big-Bang, a saber, que la densidad de las galaxias distantes debe parecerse mayor que la de las galaxias locales fue verificada por un grupo de astrónomos partidarios de la hipótesis rival, la llamada Teoría del Estado estacionario, según la cual en el Universo se está creando materia continuamente. El ejemplo está en Leplin, *op. cit.*, p. 96.

¹⁰ A veces, pues, no encontramos lo que vamos buscando, sino algo bien distinto. En 1935 H. Yukawa predijo la existencia del pi-mesón como partícula mediadora en la fuerza nuclear fuerte, por analogía con el papel desempeñado por el fotón en la fuerza electromagnética. Al año siguiente se descubrió una partícula cuya masa era ligeramente inferior a la predicha por Yukawa. Durante un tiempo los físicos pensaron equivocadamente que se trataba del pi-

Estos aspectos de la práctica científica muestran precisamente cómo una y otra vez la realidad se resiste a plegarse a nuestras expectativas e intereses. Constituyen, por tanto, contraejemplos a la tesis de que la evidencia, y consiguientemente, la noción de éxito empírico, es consciente o inconscientemente distorsionada en función de los intereses del investigador. Es verdad que en la ciencia también se producen *fraudes*. Algunos han sido descarados intentos de engaño. Otros han resultado del fervor con que los investigadores han confiado en sus ideas, lo que les ha llevado a cometer descuidos que han viciado los resultados obtenidos. Lo que nos importa aquí, de todos modos, tanto unos como otros son finalmente descubiertos. Y es que, en última instancia, la detección de un fraude no es sino una ilustración del rol arbitral desempeñado por la evidencia.

El sociologista, en consonancia con su estrategia, debería reinterpretar todos estos aspectos de la práctica científica en sus propios términos, es decir, como manifestaciones de las relaciones de influencia y poder existentes entre los distintos colectivos implicados. No me pronunciaré sobre la viabilidad de esta estrategia. Me conformo con reseñar el corolario de mi argumentación. Dado que la explicación en términos de la criba metodológica no explica por qué una teoría es empíricamente exitosa y que, además, no parece haber más alternativas al realismo aparte de la reinterpretación sociologista¹¹, entonces, si la reinterpretación sociológica de las prácticas científicas apuntadas es plausible, la explicación realista no es la mejor explicación de por qué una teoría es empíricamente exitosa; en cambio, si la reinterpretación sociológica no es plausible, la explicación realista es, por ahora, la mejor explicación.

II

Los defensores de la inferencia a la mejor explicación sostienen que, dadas varias alternativas empíricamente equivalentes, la mejor explicación es verdadera o, en su defecto, más probable que el resto. Esto es controvertido. También es objeto de controversia la posibilidad de aplicar dicha regla, aun

mesón, aunque en realidad se trataba de un mu-mesón. El pi-mesón fue descubierto en 1947. Su masa coincidía exactamente con la predicción de Yukawa. La razón de ser del mu-mesón fue un enigma durante las tres décadas posteriores.

¹¹ En "Antirealist Explanations of the Success of Science" (*Philosophy of Science. PSA Proceedings* 63 (1996), pp. S298-S305) Andre Kukla ha argumentado que el instrumentalista puede explicar el éxito empírico de una teoría T. La explicación es que el mundo observable se comporta como si T fuera verdadera (por eso las consecuencias observacionales de T son repetidamente confirmadas). Según Kukla, si se entiende esta afirmación como una verdad nomológica acerca del mundo se trata de una explicación legítima —no vacua— que no presupone el realismo. Kukla reconoce, sin embargo, que esta explicación es ininteligible (y, desde luego, que es peor que la explicación realista). Admito que resulta complicado trazar una línea de separación entre una no-explicación y una mala explicación, pero, en contra de Kukla, una explicación ininteligible me parece, sencillamente, una no-explicación.

siendo válida, al tipo de explicaciones que estamos discutiendo aquí. Pero, como dije al comienzo, no discutiré S₄. Sin embargo, llegados a este punto de la discusión, surge una cuestión más modesta, pero no menos relevante para nuestros propósitos. Concedamos que la reinterpretación sociologista no es plausible y que la verdad es la mejor explicación de por qué una teoría es empíricamente exitosa (en sentido estricto es la única explicación disponible de este fenómeno). Ahora bien, ¿es lo bastante buena en sí misma como para merecer nuestro crédito?

La irregularidad en la órbita de Urano fue atribuida por Leverrier a la influencia gravitacional de un planeta desconocido. La hipótesis newtoniana de la gravitación universal le permitió incluso calcular con bastante exactitud su masa y su órbita. Poco después, en 1846, Galle descubrió el planeta, que fue bautizado como Neptuno. En la época esto fue considerado un triunfo de la cosmología gravitacional newtoniana. Según el argumento realista, si hay éxito empírico aquí, y parece que indudablemente lo hay, deberíamos inferir la verdad de la teoría que sugiere la predicción. Mas, durante este siglo se ha demostrado convincentemente que la gravedad no es una fuerza que actúa a distancia, sino una curvatura del continuo espaciotemporal. La mecánica newtoniana es falsa. Por consiguiente, el éxito empírico no es síntoma de verdad. Así pues, el registro histórico de la ciencia amenaza con convertir la intuición realista, por básica e inapelable que resulte a primera vista, en un prejuicio, al menos cuando se trata de explotarla en el contexto que nos ocupa¹².

Circunscribir la atribución de verdad a las teorías actualmente detentadas por la ciencia, excluyendo las teorías propuestas con anterioridad conlleva cierta arbitrariedad. ¿Por qué la atribución de verdad cuenta a partir de ahora y no comenzar, por ejemplo, dentro de un siglo? Definir el éxito empírico de modo que ninguna teoría pasada pueda ser considerada propiamente exitosa¹³, lleva por idénticos derroteros. Otra sugerencia más reciente apuesta por reconocer el éxito de teorías ya arrumbadas tratando de mostrar al mismo tiempo que: (i) los componentes teóricos falsos sólo están aparentemente involucrados en la predicción, pero ésta podría haber sido llevada a cabo sin el concurso de aquéllos, y (ii) los componentes teóricos que realmente posibilitaron la predicción se conservan en la ciencia contemporánea. Con otras palabras, el aparato teórico incompatible con la teoría actual es un aditamento innecesario. La inferencia del éxito a la verdad es selectiva, esto es, funciona sólo respecto a aquellos componentes que desempeñan un papel

¹² Laudan da una lista de teorías ya abandonadas y, según él, exitosas cuyos términos centrales son referencialmente vacíos. V. *Science and Values* (Berkeley, University of California Press, 1984), p. 121.

¹³ James W. McAllister, "Scientific Realism and the Criteria for Theory-Choice", *Erkenntnis* 38 (1993), pp. 203-22.

efectivo. El realista ha de mostrar que esos componentes son, precisamente, los que han sido asimilados por la ciencia contemporánea.

Esta es una hipótesis que sólo puede contrastarse a partir de un estudio pormenorizado de ejemplos históricos conflictivos¹⁴. Sin embargo, no es necesario suscribir una concepción revolucionaria del progreso científico para admitir que cualquier disciplina científica ha sufrido notables reconceptualizaciones en los últimos ciento cincuenta años, y entonces quizá al realista le quede bien poco que salvar. Así, cuando Leplin comenta los aspectos de la concepción newtoniana a los que podemos atribuir verdad, señala que también la física actual retrotrae fenómenos como la caída de los graves, el movimiento pendular, la aceleración de los planetas, ... a un único mecanismo teórico (aunque éste sea distinto en uno y otro caso, claro está)¹⁵. Pienso, sin embargo, que ésta es una semejanza demasiado formal como para hablar de verdad o de verdad aproximada. Una atribución de verdad aproximada en este ejemplo requeriría que el mecanismo responsable de todos esos fenómenos según la teoría actual fuera una modificación o elaboración del tipo de mecanismo postulado por la teoría antigua, pero no parece que sea así. De otro modo, no será difícil encontrar algún aspecto en el que la teoría descartada y la ciencia contemporánea coincidan, vaciando así de contenido la noción de verdad aproximada.

El problema es acabar defendiendo que las teorías pasadas exitosas no eran totalmente falsas, en vez de defender que eran aproximadamente verdaderas, lo cual debilitaría la posición realista hasta el punto de hacerla trivial. No estoy diciendo que esto sea inevitable, pero resulta imprescindible para el realista una noción de verdad aproximada que, primero, no se limite a comparar las predicciones observacionales cuantitativas derivadas de las teorías en cuestión y, segundo, no defina el grado de semejanza requerido y los aspectos teóricos a tener en cuenta *a posteriori*, es decir, para casos históricos concretos¹⁶.

III

La explicación realista da cuenta de por qué una teoría es empíricamente exitosa. Dicha explicación no compite con la explicación del éxito empírico basada en la criba metodológica, ya que cada una apunta a un *explanandum* distinto. En consecuencia, determinar si la verdad es la mejor explicación del éxito empírico exige compararla con otras explicaciones, caso de

¹⁴ Vaya un botón de muestra. En *The Advancement of Science* (Oxford, Oxford U.P., 1993), cap. 5, Philip Kitcher conduce que los éxitos empíricos de la teoría de la difracción de la luz de Fresnel nada deben a su concepción del éter.

¹⁵ Leplin, *op. cit.*, pp. 148-49.

¹⁶ Al realista le queda aquí mucho por hacer. En esta línea v. *Realism Rescued*, Jerrold L. Aronson, Rom Harré y Eileen C. Way, Chicago, Open Court, 1995.

que las haya.

La reinterpretación sociologista de las nociones de éxito empírico y de evidencia bloquea el argumento realista. En ese sentido constituye una alternativa a la explicación realista, aunque hablando en propiedad no explique el éxito de una teoría. Dicha reinterpretación, no obstante, choca con la práctica científica, en concreto con aquellos aspectos que demandan un papel arbitral para la evidencia.

Mas, aunque admitamos la verdad como la mejor explicación de por qué una teoría es empíricamente exitosa, podemos plantear si se trata de una explicación lo bastante buena como para ser creída. Dada la distancia que puede llegar a haber entre ciertas teorías que fueron exitosas en el pasado y las teorías actuales que también lo son, el problema aquí es que la semejanza entre unas y otras no constituya una base suficiente para justificar la atribución de verdad aproximada. El riesgo es trivializar esta noción, y con ello, la propia tesis realista.

Como se ve, la discusión no arroja un pronunciamiento concluyente sobre el argumento realista. Espero que haya servido al menos para destacar sus puntos fuertes y sus debilidades. Ante la tesitura de un realismo trivial, tendríamos que aceptar que el éxito empírico, en el sentido que interesa al realista, es, por ahora al menos, inexplicable. Pero mientras tanto, quienes no sentimos afición por los milagros seguiremos tratando de explicarlos.

Valeriano Iranzo
e-mail: Valeriano.Iranzo@uv.es



THE TURN TOWARDS A PRIORI JUSTIFICATION

CATRIN MISSELHORN
University of Tübingen

ABSTRACT: In his recently published book *In Defense of Pure Reason* Laurence Bonjour tries to defend the possibility of synthetic a priori justification. In this article, Bonjour's position will be scrutinized, and some objections will be raised. The discussion will focus on the problem of metajustification, i.e. the problem of how to justify the truthconduciveness of a method of justification. A priori justification was traditionally supposed to have a special status in this respect, since one conceived of it as an immediate apprehension of truth. Yet, this fact as such does not render a priori justification immune to the demand for metajustification, as the examination of Bonjour's position will show. Finally, there will be made a different suggestion how to deal with the problem of metajustification concerning a priori justification.

KEYWORDS: Epistemology, justification, internalism, necessity, a priori.

I

There seems to be a tendency inherent in analytic philosophy to break more and more with some dogmas standing at the beginning of its tradition; one can even say, this development itself already belongs to the history of analytic philosophy. It began with Quine, and was continued by Davidson, McDowell, and others. In Epistemology, the possibility of synthetic a priori justification reentered the discussion after having been discredited for quite a while. This development deserves particular attention: if there is any genuine philosophical method, it is a priori justification. Nevertheless, most analytic philosophers agree that a priori justification is impossible, or restricted to the quite narrow domain of conceptual analysis. Some, most famously Quine, did not even grant that much¹. In the meantime, this dogma began to crumble². Laurence Bonjour, for instance, just published a book where he defends the possibility of synthetic a priori justification. So far, he was primarily known as an adherent of the coherence theory of justification, and his coherentist

¹ See his "Two dogmas of Empiricism" in: Quine 1953.

² See, for instance, the anthology *The Return of the A Priori* ed. by Hanson and Hunter 1992.

account was of major importance to current epistemology³. In his new book *In Defense of Pure Reason* BonJour takes an almost traditional rationalist stance on the question of a priori justification⁴.

The need of a priori justification must be seen in the light of the dialectic inherent to the epistemological debate of the last decades. One of the most vivid discussions took place between so called epistemic internalists and externalists. My hypothesis is this: the intuitions by which internalists are driven finally lead to a fall back upon the idea of a priori justification. This holds, for instance, for BonJour. Although he does not make it explicit, his turn towards the a priori is due to his internalist position⁵. Let us, therefore, rehearse what internalism claims. The main internalist intuition is that justification somehow relates to rationality. From the internalist point of view, it is not sufficient for a person's being justified in believing that p that she does - as a matter of fact - have reliable evidence for p's truth. Additionally, it must be rational from her perspective to hold it true. Hardcore externalists, on the contrary, hold that justification reduces to a reliable belief forming mechanism. In this case, p's justification may be beyond the cognitive grasp of the epistemic subject⁶. However, the divide between the two positions is not as deep as this comparison might suggest. There have been various attempts to reconcile the intuitions underlying both positions⁷. Hence, I do not want to cling too strictly to the distinction. What I do claim is that certain intuitions put forward mostly by internalists lead to problems, the idea of a priori justification is supposed to solve.

The intuitions in question are tied to the concept of justification as a rational activity, I take to be central to internalism. How can this connection be spelled out? The distinctive feature of epistemic justification is that it is aimed at truth. From a rational perspective, however, we are not only interested in having - as a matter of fact - mostly true beliefs. We want to possess good reasons for holding certain propositions true. Rationality in this sense demands, in my view, that at least a certain minimal condition is met: upon reflection the person under consideration must be able to answer any reasonable challenge to the adequacy of her evidence. If we do not restrain internalism in this way, justification will be impossible for psychological reasons. Any justification can be put into question in indefinitely many ways without good reason. We would never arrive at any justified belief if we had

³ See BonJour 1985.

⁴ See BonJour 1998.

⁵ For his commitment to internalism see BonJour 1985, 57.

⁶ BonJour's criticism of externalisms (see BonJour 1985, Chap. 3) focusses mainly on this point.

⁷ At the rather extreme ends of the spectrum range Goldman on the externalist side (see Goodman 1979) and Foley on the internalist one (see Foley 1993). Intermediate positions are proposed, for instance, by Alston (see Alston 1989) and Williams (see Williams 1996).

to reject all those unjustified challenges. This would amount to a *reductio ad absurdum* of internalism. Therefore, the weaker kind of internalism proposed is preferable, if we do not want to sever the connection between justification and rationality entirely.

According to a very prominent objection, internalism gets nevertheless involved in a regress of justification due to a demand for *metajustification*. In cases of metajustification the standards of justification themselves become the object of justification. In the next section I am going to sketch how the regress of metajustification evolves. As we will later see, the solution to this problem consists in finding a kind of justification that is not in need of metajustification any more. As BonJour seems to think, *a priori* justification can meet this demand, because it amounts to an immediate apprehension of truth⁸. This is the thesis I will scrutinize in my talk. As will turn out, BonJour does not manage to establish this claim convincingly. Therefore, I am finally going to propose a different argument in order to establish the claim that *a priori* justification is not in need of a full-blown metajustification.

II

But see first how the regress evolves. In order to meet the internalist constraint, the epistemic subject must have reasons to hold the belief in question probably true. Yet, the internalist demand of justification can be applied to these reasons, as well. Even the weak kind of internalism I propose can be challenged in this way. At least *prima facie* reasonable doubt against the adequacy of the available evidence can be raised.

Let us see how with the help of an example. Suppose, for an instant, we are adherents of a coherence theory of justification. Coherentism holds that a belief is justified if it is part of a coherent set of beliefs. According to the internalist conception of justification, a person must be able to answer any reasonable challenge of the adequacy of her evidence in order to be justified. A piece of evidence is adequate if it makes the truth of the belief under consideration probable. Now, it seems quite easy to challenge the adequacy of coherence as a method of justification. Just think of the well-known skeptical thought experiments: we could all be brains in a vat hooked up to a computer by a mad scientist, or we could be deceived by an evil demon. In both cases, our belief system could be entirely coherent. Yet, since most of the beliefs included are false, the coherence of the system would not be a good indicator for the truth of the constituent beliefs.

Thus challenged, a person must have a reason to believe that coherent belief systems are by and large true in order to be justified, i.e. a metajustification of the truthconduciveness of coherence is required. (However, let me just mention in passing that the problem of metajustification

⁸ See BonJour 1999, 131.

is independent of coherentism. It arises in a similar way for foundationalism.)⁹ The only way to avoid this consequence is to give up the internalist intuition of a fundamental tie between justification and rationality; a move that amounts to externalism. In this case, it would suffice that the beliefs contained by a coherent set of beliefs are by and large true, as a matter of fact, and there need not be any reason supporting this claim available to the subject. However, I do not want to discuss the plausibility of externalism in this context. It will be, rather, presupposed that the weak formulation of the internalist constraint proposed above captures an important intuition about justification. Let us, for the sake of argument, hold on to this intuition. If it, finally, turns out that internalism renders justification unattainable, we might still reconsider giving it up.

Yet, the prospects for the internalist account seem to be utterly bad. There is reasonable doubt whether it is possible to justify the truthconduciveness of a method of justification. Once one starts to think it through, there seems to be no way out of a regress or vicious circle: either, the method employed for metajustification is justified with the help of a different procedure which is itself in need of metajustification. Or, metajustification proceeds by using the method in question. In this case, there could not be any non-circular metajustification. In order to justify the truthconduciveness of the justification procedure we would already have to presuppose that it leads us towards the truth. There may, however, be a solution to this problem. We have to find a self-sufficient standard of justification which is not in need of metajustification. Traditionally, many philosophers have thought a priori justification could meet this demand because of its character as an unmediated rational insight into truth. We will scrutinize this claim later on. But first of all, we have to explain what a priori justification is supposed to be.

III

Commonly, a priori justification is characterized as being independent of experience. Following Bonjour, I understand experience as including "any sort of process that is perceptual in the broad sense of (a) being a causally conditioned response to particular, contingent features of the world and (b) yielding doxastic states that have as their content putative information concerning such particular, contingent features of the actual world as contrasted with other possible worlds."¹⁰ To turn it the other way round, a method of justification is independent of an appeal to experience, iff neither (a) nor (b) are true of it: a priori justification must not be a causal response to particular, contingent features of the world, and it must not yield doxastic

⁹ See Misselhorn 1999. In this talk I argued for this claim discussing another article of Bonjour's (see Bonjour 1999).

¹⁰ See Bonjour 1998, 8.

states with putative information about such features of the actual world in contrast to other possible worlds.

Yet, conditions (a) and (b) are slightly ambivalent. The first reading says that a priori justification deals with necessary truths only, i.e. propositions that are true in all possible worlds. This seems to be BonJour's view. According to the second reading, a priori justification concerns not only necessary truths, but also propositions about possible worlds which are not necessarily true, propositions of the kind "possibly p" where p is true in at least one possible world. These propositions may be justified empirically by inferring from actuality to possibility. However, they can also be justified independently of an appeal to experience, if the aim just is to show that they are true in *some* possible world which need not be the actual one. Considerations like these are very important to philosophical thought. The force of the skeptical hypotheses invoked before, for instance, depends on the thought that they are possibly true, but we are incapable of telling whether they are true in the actual world or not. Therefore, not only necessarily true propositions but also possibly true propositions can be justified a priori as long as their justification does not depend on a process fulfilling (a) and (b). In addition to this primarily negative constraint, BonJour characterizes a priori justification as rational intuition.¹¹ Rational insight somehow involves a direct and immediate grasp of the truth of a proposition. I am not sure what BonJour's claim amounts to. I myself will prefer to talk about modal intuitions, instead. However, I will use this term in a rather innocent manner, saying just that something appears metaphysically necessary or possible to us, where something is metaphysically possible iff it could be actual¹².

IV

Let us now return to the regress of metajustification. The question we have to deal with is why a priori justification is supposed to be immune to the problem. Do we not have to show that a priori justified propositions are probably true, and does not the regress, therefore, threaten again? BonJour rejects this consequence. In the case of a priori justification, he points out, the justifying feature is not an independent property the proposition possesses only for contingent reasons.¹³ In contrast, consider beliefs about the external world. We could have - at least *prima facie* - one and the same set of beliefs about the external world no matter whether they are true or false. They would not even differ in their possession of the supposed justifying property. As we

¹¹ See BonJour 1998, 131ff.

¹² Since not everything which appears metaphysically possible to us really is metaphysically possible, it is necessary to put a methodological constraint on modal intuitions. This is done with the help of the concept of conceivability. I cannot discuss this problem here, but see, for instance, Yablo 1993.

¹³ See BonJour 1998, 145ff.

saw before, a set of largely false beliefs could, for instance, be equally coherent as one constituted by mostly true beliefs. As BonJour believes, the same does not hold for a priori justified beliefs, since we hold them true only in virtue of their content.

Yet, there are two objections against this position. (1) How do we know that we grasp the content of the beliefs in question correctly? Although this is a widely discussed issue, in particular, by semantic externalists, I do not want to deal with it right here. (2) There is another more interesting argument which concerns the demand of metajustification. Suppose, we hold some propositions necessarily true only in virtue of their propositional content. Is it not possible that all this a priori evidence about necessary truths is false? Just imagine, the evil demon makes us hold a host of things necessarily true which in fact are not. Basically, the inference from a priori intuitions to metaphysical possibility faces the same problems as the inference from our conceptual structure to the empirical world¹⁴.

At this point one might feel tempted to introduce some considerations concerning semantic content. BonJour invokes a semantic theory which is rather unique in the discipline. In his view, mental content is constituted by exactly the same universals which are also instantiated by the world's facts. For this reason, it is not surprising that we can grasp necessary connections between properties by pure thinking. Yet, BonJour cannot legitimately make use of this theory in order to reject the demand of metajustification. At least not as long as he does not explain how this theory itself is justified. Either, it is justified a priori then a circle arises directly. Or, it is justified empirically. Then the circle would be somewhat larger, but no less threatening. As we argued above, the truthconduciveness of empirical justification is no less problematic than that of a priori justification. The need to establish the truthconduciveness of empirical justification even made us fall back on the idea of the a priori. There remains a last, as far as I can see, solely theoretical possibility. Besides empirical and a priori justification, there might be another method of justification. Yet, the same problem as before arises for this kind of justification, whatever it is supposed to be: we would like to know why beliefs thus justified are probably true, i.e. why the method in question is truthconducive. Finally, this argument would become circular or the regress would start anew. At best, BonJour's semantic theory can claim to be an explanation of the possibility of a priori justification. Such an explanation is, indeed, a desideratum, since there is reason to wonder how we are able to find out anything about truth by pure thinking alone. This holds at least, if we conceive of truth as a non-epistemic notion, i.e. as independent of our thought.

¹⁴ Barry Stroud raised a similar objection against transcendental arguments in general (Stroud 1968). BonJour considers something like it briefly (BonJour 1999, 149-52), but, finally, he does not recognise its importance.

But BonJour's semantic theory cannot legitimately claim to provide an adequate metajustification of the truthconduciveness of a priori justification.

V

Does this amount to saying that there is no way to avoid the threatening regress? I hope not. I will now try to develop an argument for the claim that a metajustification of a priori justification cannot reasonably be demanded. The line of reasoning proposed amounts to an indirect argument which is supposed to show that any reasonable demand for a metajustification of a priori justification does have to presuppose the truthconduciveness of this method. In this way, the regress can be stopped, and we are able to satisfy the internalist intuition.

To rehearse briefly: the truthconduciveness of a priori justification was in need of a metajustification for the following reason. Is it not possible that our modal intuitions tell us that such and such a scenario is metaphysically necessary or possible, but it is not? Well, if this is just a singular case, it does not really constitute a threat to a priori justification, since we need not claim that it is infallible. However, if the major part of our modal intuitions could be false, it would undermine the truthconduciveness of a priori justification. In this case, it would not be a good indicator of the truth of the belief in question. It seems to be quite easy to imagine such a situation, for instance, in a demon world where all of our beliefs about necessary truth are false. Therefore, this objection amounts *prima facie* to a reasonable challenge of the method of a priori justification. Yet, there is a little detail about this objection we should direct our attention at. It contains the modal formulation "is it not possible that." This might suggest that the objection itself is based on modal intuitions, and, therefore, presupposes their truthconduciveness.

Yet, a skeptic about a priori justification might not agree with this consequence. He might simply deny that the challenge to the truthconduciveness of a priori justification relies on modal intuitions. Instead, he might argue, it suffices to show that most things considered to be necessarily true in the history of philosophy and the history of the sciences turned out to be false, as a matter of fact. In this way, one could undermine the truthconduciveness of a priori justification at least with respect to necessary truths by empirical evidence. Yet, this is a rather weak objection which will depend on weighing the quantity and the importance of cases, the reasons for errors, and so on. Moreover, many famous counterexamples from the history of the sciences (think of Euclidean Geometry, for instance) might not have been justified a priori, at all.

However, there is another more important objection against the empirical strategy to undermine a priori justification. Given internalism, we cannot simply presuppose that our empirical beliefs are justified. Rather, we had to fall back on a priori justification in order to justify the

truthconduciveness of empirical justification. Therefore, if we have reason to doubt the possibility of a priori justification this is all the worse for our empirical knowledge. We seem to face a dilemma here: either, we claim to have empirical knowledge, then we nolens volens also have to accept that a priori justification is truthconducive. Or, we lack empirical knowledge in which case it cannot undermine a priori justification.

Yet, not just the argument for the truthconduciveness of empirical justification has to be a priori. Modal intuitions enter the picture even if one tries to attack the reliability of empirical justification. As we saw above, the skeptical challenge depends on the possibility of skeptical hypotheses like the brains-in-the vat scenario, or Descarte's evil demon deceiving us. Obviously, a reasonable challenge consists in more than just the claim that all of our beliefs of a certain type could be false. This possibility seems to be in need of a rather detailed elaboration in order to constitute a serious threat to our knowledge claims. The point of the hypotheses is, therefore, to provide alternative explanations of our evidence which make the beliefs supported by the evidence largely false. If skeptical hypotheses could be true, our confidence in the truthconduciveness of our methods of justification would be unjustified, since they could provide us with mostly false beliefs. However, this involves modal claims relying on modal intuitions of possible worlds furnished with evil demons, mad scientists and other fancy stuff. Therefore, skepticism presupposes the truthconduciveness of a priori justification. In other words, the skeptic has to be confident that what seems metaphysically possible to him really is.

Yet, at this point the skeptic might dig in the heels. He could insist that skeptical possibilities do not have to be metaphysical possibilities in order to undermine our knowledge claims. Rather, it would suffice, if they were just logically possible, i.e. logically consistent. And the logical consistency of skeptical hypotheses can hardly be denied if one wants to hold on to a realist conception of truth and a fallibilist conception of justification. A minimal realism says, to quote Michael Williams, that "thinking does not make it so"¹⁵. Or, to put it in another way, an objective proposition's being true is one thing, and our believing it to be true, something else. If this is right, there always exists a logical gap between our beliefs and the world, since our evidence cannot logically guarantee the truth of the beliefs it is supposed to support. For the same reason, fallibilism holds, i.e. any of our justified beliefs may be false. But exactly that makes for the problem. If the connection of realism and fallibilism just amounts to the logical possibility of skepticism, it cannot be enough to undermine our knowledge claims. If fallibilism would suffice to generate skepticism, a fallibilist conception of knowledge would be a conceptual impossibility! A position nowadays almost nobody is prepared to

¹⁵ See Williams 1996, 228.

hold¹⁶. For this reason, the logical possibility of skeptical hypotheses is not at all threatening.

To arrive at skeptical conclusions, skeptical hypotheses have to claim the status of metaphysical possibilities. As such, their justification has to rely on modal intuitions which are, in my view, a priori. Finally, skepticism about a priori justification turns out to be an epistemically inconsistent position. Its epistemic inconsistency lies in the fact that the truth of skeptical hypotheses with respect to modal intuitions would undermine their justification. Despite appearances, the skeptical challenge of the a priori does not seem to be a reasonable one, after all. Even philosophical skepticism is confined to the bounds of reason. To show this much is sufficient to deal with it within the internalist framework proposed. If this is correct, it is possible to hold on to the internalist intuition of a fundamental tie between rationality and justification without rendering justification unattainable.

To conclude a rather dry essay with a little more literary glamour: in some sense, it does not altogether come as a surprise that a priori justification turns out to be at the bottom of epistemology. It seems to be an unavoidable presupposition of any philosophical scrutiny that, in the end, things will turn out to be as reason tells us they must be.

REFERENCES

- Alston, W.P. (1989): *Epistemic Justification. Essays in the Theory of Knowledge*, Ithaca.
- BonJour, L.: (1985): *The Structure of Empirical Knowledge*, Cambridge/MA.
- (1998): *In Defense of Pure Reason*, Cambridge.
- (1999): "The Dialectic of Foundationalism and Coherentism", in: *The Blackwell Guide to Epistemology*, ed. by J. Greco/E. Sosa, Malden/MA; Oxford.
- Foley, R. (1993): *Working Without a Net. A Study of Egocentric Epistemology*, New York; Oxford.
- Goldman, A.: (1979): "What Is Justified Belief?" in: *Justification and Knowledge*, ed. by G. Pappas, Dordrecht.
- Hanson, P./Hunter, B. (1992): *The Return of the A Priori*, Calgary.
- Misselhorn, C. (1999): "Internalism and the Dialectic of Foundationalism and Coherentism", Talk given at the conference *Challenges to Traditional Epistemology*, 14.-16.1.99 in Tübingen (German translation in preparation).
- Quine, W. V. O. (1953): "Two Dogmas of Empiricism", in: *From a Logical Point of View*, 2nd, revised Edition, Cambridge/MA; London/England 1980.
- Stroud, B.: (1968) "Transcendental Arguments", in: *The Journal of Philosophy* Vol. LXV, No. 9.
- (1989) "Understanding Human Knowledge in General", in: *Knowledge and Scepticism*, ed. by M. Clay/K. Lehrer, Boulder.
- Williams, M. (1996): *Unnatural Doubts. Epistemological Realism and the Basis of Skepticism*, Princeton/NJ.

¹⁶ See, for instance, Stroud 1989, 35.

Catrin Misselhorn

University of Tübingen

e-mail: catrin.misselhorn@student.uni-tuebingen.de



NUEVAS TENDENCIAS EN CIENCIAS COGNITIVAS Y FILOSOFÍA NATURALIZADA DE LA CIENCIA

ALVARO MORENO

Unidad Asociada de Filosofía de la Ciencia (UPV/EHU-CSIC/Instituto de Filosofía)

ABSTRACT: Recent theories conceive cognition intrinsically related with some functional action in the environment (and not merely a mirror reflect of reality). This view of cognition as an interactive process able to sustain the functional viability of highly complex systems, implies that there are many more factors which play a role in the study of cognition than what used to be considered in classical theories. This new perspective on cognition affects scientific knowledge as it is a very special kind of cognitive activity.

KEYWORDS: Naturalization, theory of evolution, cognition, science

1. Introducción

En su concepción tradicional, el objeto de la filosofía de la ciencia es el estudio del conocimiento científico, concebido como un metanivel ajeno a toda problemática relacionada con los contenidos de las ciencias empíricas. La filosofía, pues, pertenece al nivel epistemológico, y éste se ha considerado completamente separado del ontológico. El proyecto de naturalización concibe en cambio el conocimiento como un fenómeno natural producido en el curso de la evolución biológica. Por tanto, susceptible de ser estudiado por las propias ciencias empíricas (biología, neurociencia, psicología cognitiva, etc).

La propuesta de naturalización se basa en la idea de que todos los aspectos de la realidad -lo físico, lo biológico, lo psíquico, lo social- forman parte de un único mundo natural. El mundo es pues una realidad integral y no existen localmente sustancias específicas. El naturalismo implica un rechazo fundamental a cualquier tipo de ontología, postulado o explicación ad hoc, más allá de los cuales no quepa formular cuestiones. Por consiguiente, también los procesos cognitivos forman parte del mundo natural y pueden (y deben) ser explicados en base a las teorías científicas.

Aunque una de las fuentes de la idea de naturalización del conocimiento son los trabajos de la llamada Escuela de Viena (Lorenz, Vollmer, Riedl...), en el campo de la filosofía de la ciencia fue Quine (1959) quien, hace ahora treinta años, formuló la propuesta de “naturalizar” la epistemología al sostener que el núcleo más específicamente filosófico de la filosofía de la ciencia de la Concepción Heredada está contenida en la ciencia natural. En las dos últimas décadas, las ideas naturalizadoras han obtenido una notable repercusión en el ámbito de la filosofía de la ciencia, quizás, como señala A. Rosenberg (1996), porque parecen ofrecer una alternativa tanto a las escuelas de la filosofía de la ciencia que parten en su análisis de un (meta)nivel epistemológicamente previo al conocimiento como al relativismo de los constructivistas sociales. Con su propuesta de sustituir la epistemología como metanivel fundante de la filosofía de la ciencia por teorías científicas empíricamente testables, el naturalismo busca reconciliar la historia de la ciencia como resultado de una acción social con la visión positivista de la ciencia como un sistema productor de conocimiento racional y objetivo, independiente en última instancia, de su contexto psicológico y social.

2. Naturalización, Teoría de la Evolución y Cognición

Sin embargo, la utilidad que la naturalización puede representar para la filosofía de la ciencia -y, en última instancia, para la auto-comprensión de la ciencia- pasa a mi entender, por un esfuerzo de interpretación crítico de los contenidos de las ciencias biológicas¹ y cognitivas que vayan a ser utilizados como base de la naturalización del conocimiento.

De hecho, algunas de las críticas más contundentes que se han formulado contra las propuestas naturalistas lo son más contra formas específicas del naturalismo que contra su propuesta conceptual básica. Por ejemplo, Thagard (1980) ha argumentado que la pretensión de explicar mediante los principios de la Teoría de la Evolución neodarwiniana los mecanismos de la evolución del conocimiento científico resulta insostenible. Y Hooker (1995) ha sostenido también que el proyecto de naturalización del conocimiento requiere una reformulación tanto de la teoría clásica de la evolución como de ciertos principios de la IA clásica y conexionista.

Vamos pues a analizar brevemente la raíz de estas interpretaciones erróneas.

¹ La biología -y de forma particular, la teoría darwiniana de la evolución- juega un papel muy importante en muchos de los planteamientos modernos de naturalización (Kitcher 1993, Hull 1988, Callebaut 1993, Ruse 1995, Campbell 74). Pero el planteamiento que hacemos aquí parte de una visión más amplia de la biología que la del evolucionismo darwiniano ortodoxo.

1. La Teoría clásica de la evolución ha considerado que el desarrollo de las capacidades cognitivas en los animales (incluido el hombre) es una forma de adaptación bajo la presión de la Selección Natural. En esta concepción la formación de estructuras y comportamientos cognitivos se ve como una adaptación a una realidad externa pre-existente. Sin embargo, las estructuras cognitivas de los organismos que resultan viables se han construido en el marco de un proceso de mutua co-evolución con su propio entorno. El conjunto de objetos y eventos de tales sistemas cognitivos (la "realidad") está formado en su casi totalidad por otros organismos cuya estructura ha coevolucionado -y sigue haciéndolo- con la del sujeto. El concepto de la relación organismo-entorno, desde el punto de vista de la evolución, más que de una relación adaptativa concebida como un proceso activo por una parte (la especie X) y pasivo por otra (el entorno Y), debe ser pensado como una relación simétrica, de ajuste mutuo, en definitiva, como un proceso co-evolutivo. La adaptación es una selección mutua de organismos y entornos. Por lo tanto, en gran medida, todo el sistema sujeto-objeto es fruto de un proceso de co-constitución. La evolución no es pues un conjunto de procesos de adaptación de los organismos a un entorno previo e independiente de éstos, sino un conjunto de procesos de ajuste mutuo entre una determinada población de organismos y un entorno dinámico, constituido fundamentalmente por otros (tipos de) organismos (e incluso, en parte, el propio entorno físico es a su vez modificado por los organismos que se "adaptan" a él).

Pero si ontológicamente hablando agentes cognitivos y entornos se constituyen mutuamente, epistémicamente puede hablarse (a través de un proceso de corrección de errores, por ej) de objetividad en las relaciones cognitivas. Lo que ocurre es que las relaciones causales en uno y otro caso ocurren en procesos temporales diferentes.

2. El seleccionismo neodarwinista tiende a ignorar el papel que juegan los procesos de interacción con el entorno en el desarrollo de los fenotipos reales. Ello conduce a una errónea concepción (de hecho, una sobreestimación) del papel jugado por los genes en la configuración de la estructura corporal y, por consiguiente, a no comprender adecuadamente tampoco los procesos de configuración e inserción de los sistemas cognitivos en el conjunto de la estructura corporal y la dinámica de sus interacciones con el entorno. Se trata en realidad de un proceso donde diversos niveles jerárquicos modulan los mecanismos de desarrollo ontogenético, y donde, a su vez, las interacciones dinámicas con el entorno constriñen el espacio de las configuraciones fenotípicamente viables. Por consiguiente, parece claro que la evolución de las

estrategias cognitivas debe ser explicada en función de los mecanismos que gobiernan los procesos constructivos internos en el marco del establecimiento (o logro) de ciclos interactivos con el entorno viables para los individuos y, sobre todo, para las especies.

Así pues, toda una serie de conceptos sobre la representación como proceso de adecuación y sobre la irrelevancia de las interacciones a escala ontogenética en los procesos de configuración de las estructuras cognitivas, que se han solido presentar en muchas propuestas de naturalización (especialmente en el caso de la Epistemología Evolucionista) responden no tanto a los datos de las ciencias empíricas, sino más bien a interpretaciones demasiado simplificadas de ciertas teorías²

Con todo, estas observaciones críticas son menos importantes que la propia carencia de una teoría explicativa del surgimiento y función específica del fenómeno cognitivo. A mi juicio, todo proyecto de naturalización del conocimiento en general debe empezar por deslindar lo cognitivo de su *background* biológico y proporcionar una explicación fundacional de lo cognitivo. Esta clave fundacional de lo cognitivo reside, como veremos en el punto siguiente, en la aparición de un sistema que permitió la autonomización de unos mecanismos constructivos internos sobre base informacional del resto del mantenimiento energético-constructivo global del organismo, al tiempo que dicho sistema, en su funcionalidad global, estaba plenamente corporeizado.

¿Por qué es importante partir de una explicación fundacional de lo cognitivo? Porque ello tiene tres consecuencias importantes:

1) A nivel filogenético: En vez de ver únicamente la evolución de la cognición como un resultado de los principios generales de la evolución biológica, en realidad la emergencia del fenómeno cognitivo cambia dramáticamente al propio proceso evolutivo, no en el sentido de modificar sus mecanismos básicos, pero sí en el de causar un incremento cualitativo en la complejidad de sus resultados.

2) A nivel ontogenético, al producir también un incremento cualitativo en la plasticidad y complejidad de los procesos adaptativos de los organismos (fundamentalmente, en forma de nuevas capacidades y comportamientos: memoria, aprendizaje, comportamiento anticipatorio, cooperación social, uso de instrumentos...)

3) Finalmente, todo ello lleva a la propia superación de la evolución propiamente biológica por la cultural.

² En este sentido, una consideración que me parece importante es la necesidad de adoptar en estas cuestiones un enfoque sistemáticamente interdisciplinario.

Por consiguiente, la naturalización de la cognición en base exclusivamente de la teoría de la evolución no nos proporciona los elementos conceptuales necesarios para abordar con posibilidades de éxito un proyecto de naturalización de las formas más complejas y evolucionadas de la cognición.

A partir de estas observaciones críticas vamos a replantear el concepto de cognición desde sus orígenes biológicos y, a continuación, trataremos de formular los principios que han guiado su evolución hasta llegar a su expresión más desarrollada y compleja -la ciencia- y mostrar que el conocimiento mismo y las condiciones que lo generan no deben ser considerados como dos cuestiones separadas.

3. La cognición desde el marco biológico

El origen y desarrollo de la capacidad cognitiva en la Naturaleza están ligados a la evolución de organismos pluricelulares cuya estrategia de vida dependía del movimiento. En aquellas formas de vida basadas en la búsqueda de alimento a través del movimiento se generó un proceso de diferenciación celular que llevó a la formación de un subsistema interno capaz de conectar rápida y plásticamente las superficies efectoras y sensoras. Este proceso fue el origen del sistema nervioso.

Si las posibles respuestas adaptativas estuvieran basadas exclusivamente en la versatilidad metabólica del organismo (como ocurre en unicelulares y plantas), la complejidad y rapidez del repertorio de comportamientos estaría seriamente limitada. Esta es la razón por la que sin la aparición de un subsistema interno tal como el nervioso no es posible proporcionar, en organismos pluricelulares, plasticidad y rapidez en el comportamiento adaptativo, y por consiguiente, tampoco es posible la evolución hacia lo que entendemos por comportamiento cognitivo.

Las operaciones del sistema nervioso implicaron el desarrollo de un universo interno de configuraciones con una referencialidad externa (debido a su vinculación con los sensores y efectores) organizadas en una red auto-modificante. Así, en los organismos dotados de sistema nervioso (los animales) la adaptación se produce no a través de mecanismos metabólicos de auto-control, sino de un meta-control informacional sobre las funciones metabólico-motoras. Esta es la razón por la que con el desarrollo del sistema nervioso, los animales han sido capaces de producir e interpretar configuraciones informacionales externas (signos).

Así pues, podemos definir la cognición como una función biológica de adaptación a nivel ontogenético, que desarrolla en el interior del organismo un universo plástico (y potencialmente ilimitado) de configuraciones

informacionales intencionales (representaciones), operacional (Maturana y Varela, 1990) y energéticamente autónomas respecto del resto de las funciones biológicas básicas (metabolismo), pero funcionalmente integradas en el conjunto del organismo³ (Moreno et al, 97).

A lo largo de la evolución animal las capacidades cognitivas se han ido complejizando de diversas formas: distribuyéndose en organizaciones sociales (a través de procedimientos de comunicación), generando nuevas formas de recursividad anticipativa (como el desarrollo de nuevas formas de control y monitorización autodirigidas de aprendizaje), permitiendo la extensión de las capacidades corporales mediante el uso de instrumentos, etc. Pero ha sido en la especie humana donde todo el proceso evolutivo anterior ha generado nuevas formas cualitativamente más complejas, como el lenguaje simbólico, el uso sistemático de medios técnicos, la escritura, la organización social de sistemas de detección de errores cada vez más intersubjetivos y descontextualizados, etc., dando lugar así al conocimiento propiamente racional

La evolución tan acelerada de la función epistémica en el hombre (evolución ya cultural y técnica, no propiamente biológica) ha ido creando un sistema de procedimientos colectivos de aprendizaje cada vez más universales y racionales (matematización, metodologías experimentales despersonalizadas, etc) que han generado a su vez la imagen de un correlato referencial objetivo de tales procesos. El resultado del desarrollo combinado de todos estos factores ha sido la ciencia.

4. La ciencia desde la perspectiva naturalista

La ciencia es una forma particularmente desarrollada de conocimiento racional, basada en procedimientos de evaluación empírica colectivamente acordados y mediados tecnológicamente. Como han señalado Campbell (1997) y Christensen y Hooker (1998), la ciencia es un proceso de aprendizaje autodirigido colectivamente organizado.

Pero, como hemos dicho más arriba, la capacidad cognitiva está, desde su origen biológico, esencialmente ligada a la acción. Lo cognitivo es inseparable de la interacción funcional con el entorno, pues la cognición surge como una forma de acción adaptativa y no desligable de la corporalidad. De hecho, la cognición es un proceso interactivo y constructivo de constricciones que organizan y dirigen de forma cada vez más compleja las interacciones

³ Por el contrario, la tendencia dominante en las Ciencias Cognitivas (y especialmente en la IA) es la de plantear las funciones cognitivas al margen o separadamente de las funciones biológicas básicas.

cognitivas, como aprendizaje, anticipación, etc. Debemos, por consiguiente, ver el proceso del desarrollo de la ciencia como un proceso dinámico y colectivo que hunde sus raíces en el desarrollo de las funcionalidades adaptativas y cognitivas a lo largo de la evolución biológica.

Desde una perspectiva naturalista la visión de la ciencia como acción frente a la concepción propiamente epistémica (Hacking 1983) no aparece como una contradicción. Al contrario, si analizamos los factores que han originado y gobernado la evolución de la función cognitiva, veremos que el desarrollo de las representaciones epistémicas (representaciones más o menos “filtradas” por *a priori*s diversos) resulta inseparable de la acción.

Así pues, la ciencia no es únicamente un proceso de construcción de modelos o estructuras conceptuales con capacidades predictivas referidas al mundo real, sino también la modificación profunda de este mismo mundo real por dichos modelos y para la propia construcción de éstos. La ciencia, entendida así como una actividad -basada en el conocimiento- que transforma la realidad natural y social, resulta indisociable de la técnica.

Pero la ciencia no ha inventado la tecnología. Las técnicas son tan antiguas como el hombre. Se ha argumentado incluso (Leroi-Gourhan, 1964, Stiegler, 1994, 1996) que el desarrollo de técnicas es la peculiaridad del estatus biológico del hombre, que lo ha constituido, porque la invención continuada de nuevas técnicas ha actuado como un continuo desarrollo de prótesis de los órganos sensoriales, motores, de procesamiento y almacenamiento de información, etc. B. Stiegler (1999) ha denominado a este proceso “epifilogénesis” porque es una especie de evolución de nuevas capacidades biológicas sobre una permanencia de su estructura básica. La tecno-ciencia es considerada por este autor como una exo-somatización constitutiva de lo humano. Por consiguiente, la novedad que supone la invención de la ciencia es la imbricación radical entre conocimiento racional y tecnología. En este sentido, podemos considerar a la actividad científica como la forma más refinada y sistemáticamente organizada de ese proceso de tecno-constitución del hombre.

Esta acción tecno-cognitiva racionalmente dirigida que es la ciencia va inevitablemente ligada a una manipulación a gran escala no sólo de sus objetos de estudio, sino también de la organización social de los sujetos que los estudian. La ciencia constituye pues una actividad cognitiva, organizada colectiva y racionalmente, imbricada en la actividad tecnológica y basada en procedimientos colectivamente estipulados, organizados e institucionalizados de producción, transmisión y revisión de sus contenidos y metodologías. Por eso la ciencia es una forma de conocimiento inseparablemente unida a un

conjunto de prácticas; estas prácticas producen el conocimiento científico, y a su vez, recursivamente, dicho conocimiento configura las acciones prácticas que lo posibilitan⁴. Tal vez lo verdaderamente específico de la ciencia sea precisamente la integración de actividades altamente teóricas, con otras técnicas y sociales, en un proceso global y coherente. La ciencia es, pues, un sistema formado por diferentes niveles imbricados, donde hasta sus propios productos pueden llegar a jugar un papel en el nivel que gobierna sus condiciones de producción; en definitiva, en ayudar a modificar consciente y racionalmente su propio funcionamiento cognitivo.

5. Consecuencias para la naturalización de la filosofía de la ciencia

Las diversas filosofías de la ciencia de tendencia formalista han intentado explicar la dinámica de las teorías científicas mediante un conjunto de reglas formales independientes en lo esencial del fenómeno social y material que constituye la ciencia real. Pero, a pesar de su creciente complejidad y sofisticación, siguen sin resolver el problema central de que muchas de las más importantes decisiones científicas no han podido ser explicadas en términos de cualesquiera criterios de racionalidad propuestos por dichas filosofías.

En realidad, en todas esas filosofías hay una separación muy difícil de franquear entre la reflexión sobre la ciencia como instrumento de conocimiento (la filosofía como investigación sobre los principios racionales del método científico entendiéndolo como un modelo ideal de racionalidad) y el estudio de las complejas condiciones históricas, sociales, psicológicas, etc. en que se produce y evoluciona la ciencia. Una de las razones de la crisis sin término de la Concepción Heredada (y de sus sucesoras) reside en la incapacidad de dar cuenta de la tensión entre la complejidad del sistema cognitivo tecnico-científico y la aparente simplicidad de la mayor parte de sus resultados. De hecho, la progresiva toma de conciencia de que el proceso real del desarrollo científico era mucho más complejo de como se había planteado, ha llevado a muchos autores a renunciar a todo intento de salvar el principio de reconstrucción racional. Por eso, sobre estas bases no es posible establecer la relación a la que aludíamos en los puntos anteriores entre las implicaciones de los productos del conocimiento científico y las de las teorías sobre la acción y cognición humanas para la producción de tales conocimientos. Y por eso, una nueva filosofía de la ciencia naturalizada entendida como un estudio integral de la ciencia (que englobe el análisis de la estructura conceptual de la ciencia, el de sus contenidos desde una perspectiva ontológica y el análisis de la actividad

⁴ En realidad, este esquema básico existe ya en los procesos de aprendizaje animal.

científica en tanto que proceso tecnico-social) debe hacerse a través de una teoría sobre los sistemas complejos.

A diferencia de los modelos lógicos del método científico en las filosofías de la ciencia de orientación formalista, que ignoran la interacción con el entorno natural y social en el desarrollo de la ciencia, esta propuesta de naturalización de la filosofía de la ciencia permite analizar el fenómeno científico no sólo teniendo en cuenta sus diferentes dimensiones, sino en continuidad e interacción con los contenidos de las ciencias naturales. Esta propuesta permite también integrar tanto la dimensión social como la natural en la explicación del proceso científico -esto es, la interacción colectiva de los científicos con su entorno social y natural- pues concibe el fenómeno científico (incluyendo su dimensión social) dinámicamente, y es desde esta visión como lo inserta en el marco más amplio de la evolución biológica. La filosofía naturalizada concibe pues a la ciencia (por muy sofisticada y abstracta que se nos aparezca) como una actividad cognitiva que en definitiva hunde sus raíces en los mecanismos biológicos de la cognición, y por tanto debe ser explicada mediante principios que sean coherentes con las propias teorías científicas sobre el origen y la evolución de la cognición animal.

Así, dos de los fundamentos de la filosofía clásica de la ciencia -el empirismo y la neutralidad ontológica del formalismo- resultan radicalmente cuestionados desde la perspectiva de la naturalización. Por un lado, porque la pretensión de fundamentar la ciencia sobre una base empírica (esto es, en última instancia, sobre nuestra experiencia sensorial) termina revelándose inconsistente cuando las modernas teorías sobre la naturaleza de los procesos cognitivos humanos y animales ponen de manifiesto el fundamental papel que juegan numerosos y complejos niveles de constricciones en la estructuración de la información sensorial. Y por otro lado, porque parece difícil negar que las constricciones más universales en el tratamiento de la información responden en realidad a formas muy comunes de la interacción biológico-cognitiva de los organismos con una gran variedad de entornos, que se han seleccionado. En este sentido, las estructuras más profundas y universales evolutivamente hablando del sistema nervioso sí codificarían una cierta información sobre la "realidad".

Esto significa que las teorías de la filosofía de la ciencia deben ser coherentes con el proceso evolutivo de la cognición, de manera que se muestre cómo sus desarrollos funcionales progresivamente complejos han ido prefigurando la estructura general del conocimiento científico. Apoyándose en investigaciones sectoriales, una filosofía de la ciencia naturalizada debe tratar de reconstruir los mecanismos que guían la evolución del conocimiento

científico a partir de los de la cognición animal, desde sus orígenes. Esto implica, tratar las categorías normativas relevantes de la filosofía de la ciencia (como racionalidad y criterios epistémicos) en términos tales que permitan reconstruir un sistema normativo dentro de dicho marco dinámico. Sin pretensiones reduccionistas, el objetivo es mostrar cómo a partir de los mecanismos fundamentales de la cognición animal se pueden producir estructuras y formas de interacción epistémico-funcionales con el entorno cada vez más complejas; y cómo las complejas constricciones que se van generando en este largo proceso evolutivo constituyen el origen de los principios que regulan y gestionan los diferentes tipos de interacciones epistémicas, desde los métodos hasta las instituciones científicas, pasando por las teorías, las metafísicas, las tecnologías y las prácticas experimentales (Hooker y Christensen 98).

6. Conclusión

El conocimiento mismo (los contenidos de éste o representaciones del mundo) y las condiciones que lo generan no deben ser considerados como dos cuestiones separables.

Si consideramos a la ciencia no sólo como conocimiento, sino también como actividad, y, sobre todo, si consideramos que el conocimiento generado no puede ser comprendido sin analizar la actividad que lo genera (las condiciones de su producción) y las consecuencias teóricas y prácticas que entraña, la filosofía de la ciencia naturalizada debe tratar de formular una teoría sobre el fenómeno científico en continuidad con la actividad cognitiva, considerándola desde sus fundamentos biológicos, es decir, en continuidad e interacción con los contenidos de las ciencias naturales.

Este planteamiento supone una cierta relación de circularidad entre las implicaciones de los productos del conocimiento científico (que fundamentan y guían las teorías sobre la cognición) y las implicaciones de las teorías sobre la acción y cognición humanas para la producción de dichos conocimientos. Por eso, en realidad una filosofía de la ciencia concebida de esta manera implica un programa de investigación que trate de compaginar los contenidos de las teorías científicas con la justificación epistemológica de sus condiciones de producción.

La piedra de toque de toda teoría sobre la ciencia con pretensión globalizadora será entonces el grado de coherencia que mantenga con las teorías sobre la fundamentación biológica de la cognición. Esto significa que las teorías de la filosofía de la ciencia deben ser coherentes con la explicación del proceso evolutivo de la cognición, de manera que se muestre cómo sus desarrollos

funcionales progresivamente complejos han ido prefigurando la estructura general del conocimiento científico.

BIBLIOGRAFÍA

- Callebaut, W.(1993): *Taking the Naturalist Turn or How Real Science is Done* Chicago University Press.
- Campbell, D. T. (1974), "Evolutionary epistemology" en Schilpp, P. A. (ed.) (1974), *The philosophy of Karl Popper*. (La Salle: Open Court)
- Campbell, D.T. (1997) "From evolutionary epistemology via selection theory to a sociology of scientific validity", *Evolution and Cognition*, **3**, 5-38.
- Christensen, W.D. y Hooker, C.A. (1998). "From Cell to Scientist:Toward an Organisational Theory of Life and Mind", en J. Bigelow (ed) *Proceedings of the Australian Academy of Humanities*. (En prensa).
- Gould, S.J. & Lewontin, R.C. [1979], 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme', *Proceedings of the Royal Society of London* **205**, 581-598.
- Hooker, C.A. [1995], *Reason, Regulation and Realism: Toward a Naturalistic, Regulatory Systems Theory of Reason*. (Albany, N.Y.: State University of New York Press.)
- Hacking, I. (1983) *Representing and intervening* Cambridge University Press.(Traducción cast. Paidós México)
- Hull, D. (1988) *Science as a Process* Chicago University Press
- Kitcher, P (1993): *The Advancement of Science* Oxford University Press
- Leroi-Gourhan, A.(1964): *Le geste et la parole*, Paris, Albin Michel.
- Maturana, H y Varela, F. (1990): *El árbol del conocimiento*, Barcelona: Debate
- Moreno, A, Umerez, J e Ibañez, J. (1997):"Cognition and Life. The Autonomy of Cognition". *Brain & Cognition* **34 (1)** pp. 107-129
- Quine, W. (1969): "Epistemology Naturalized" En *Ontological Relativity and Other Essays*. Columbia University Press.
- Rosenberg, A. (1996): "A Field Guide to Recent Species of Naturalism". *Brit. J. Phil Science* **47** pp 1-29
- Ruse, M. (1995): *Evolutionary Naturalism* Routledge
- Stiegler, B. (1994, 1996) *La technique et le temps*, tome 1, *La faute d'Epiméthée*, y tomo 2 *La Désorientation..* Editions Galilée, Paris.
- Stiegler, B, (1999) "La finitude rétionnelle", en *Mémoire collective et technique*, Numero especial de *Technologies, idéologies, pratiques* .(En prensa).
- Thagard, P. (1980): "Against Evolutionary Epistemology" en P.D.Asquith y R.N.Giere (eds) *PSA 1980*. pp 187-196

Alvaro Moreno Bergareche
Universidad del Pais Vasco UPV/EHU
e-mail: ylpmobea@sf.ehu.es



NATURALISMO EN FILOSOFÍA DE LA CIENCIA: Pasado y futuro

F. JAVIER RODRÍGUEZ ALCÁZAR
Universidad de Granada

ABSTRACT: Naturalism in epistemology and philosophy of science has been often criticized on the grounds that it surrenders the normative tasks traditionally assigned to those disciplines. After offering a reminder of how this objection is addressed in the *locum classicum* of contemporary naturalism (Quine's "naturalized epistemology") I argue that the crucial test for assessing the normative performance of a naturalistic stance lies in its account for the goals (or values) of science. Then I compare R. Giere's, P. Kitcher's and L. Laudan's accounts, pointing to their respective shortcomings and introducing the main features of an alternative.

KEYWORDS: Naturalism, goals, values, Quine, Giere, Kitcher, Laudan.

1. El naturalismo y la filosofía analítica

La tradición analítica dio sus primeros pasos marcada, en la obra de Frege, por la reacción contra una forma de naturalismo, el psicologismo en la filosofía de la lógica. Poco después, otro filósofo analítico, G.E. Moore, atacaba otra versión del naturalismo, el ético, con un argumento (el de la *open question*), que puede considerarse paradigmático del estilo analítico de hacer filosofía. Paralelamente, el movimiento fenomenológico y, con él, la filosofía "continental" contemporánea arrancaban con un pensamiento, el de Husserl, también radicalmente anti-psicologista. Así pues, a comienzos del siglo que está acabando, el naturalismo ofrecía, a uno y otro lado del Canal de la Mancha, el aspecto de una plaga a punto de ser exterminada. Sin embargo, cuando ese mismo siglo está cercano a su final, la polémica en torno al naturalismo sigue, para irritación de muchos, más viva que nunca. No sólo eso. Algunos llevamos la provocación hasta el extremo de sugerir que el naturalismo es una de las corrientes con más porvenir dentro de eso a lo que, con los términos cada vez más cargados de equivocidad, seguimos refiriéndonos como "filosofía analítica". En este trabajo voy a reivindicar un lugar para el naturalismo en de una de las disciplinas que tradicionalmente

han recibido más atención dentro de la tradición analítica: la filosofía de la ciencia. Sin embargo, al mismo tiempo sostendré que esa contribución futura exige algunas correcciones en las estrategias naturalistas con más predicamento. Me concentraré en las propuestas de tres autores (Larry Laudan, Ronald Giere y Philip Kitcher) y en un problema que considero auténtica piedra de toque de una posición naturalista: su tratamiento de los *objetivos* de la ciencia.

El naturalismo reivindicó la pertinencia para la epistemología de la investigación empírica sobre el conocimiento. Pero, al hacerlo, planteó una importante duda con respecto a las funciones de esa disciplina. ¿Podía una epistemología “naturalizada” seguir desempeñando las tareas propias de la teoría del conocimiento? Evidentemente, la respuesta a esta pregunta depende de qué tareas se consideren *propias* de la epistemología y cuáles se asimilen a los objetivos espurios de alguna tradición epistemológica particular. El principal representante del naturalismo epistemológico a partir de los años cincuenta, W.V. Quine, aceptó gustosamente el abandono de una de las pretensiones de la epistemología tradicional: la de *fundamentar* el conocimiento científico. Sin embargo, reclamó para la epistemología naturalizada la pretensión más genérica de *justificarlo*. De hecho, ésta ha sido y sigue siendo, quizás, la principal promesa del naturalismo: una concepción de la justificación que proporcione una anhelada vía intermedia entre el caduco fundamentismo y el inquietante relativismo¹. De esta forma, Quine se empeñaba en asumir, contra lo supuesto por muchos de sus críticos², la tradicional dimensión *normativa* de la epistemología, que se concreta tanto en la justificación en términos generales de la posibilidad de genuino conocimiento (respuesta al escéptico) como en el desarrollo de una teoría de la evidencia que dé cuenta de la justificación de secciones particulares del edificio cognoscitivo (de forma destacada, las teorías científicas). Ahora bien, si la vocación normativa de la epistemología naturalizada debería quedar fuera de toda sospecha, está mucho menos claro que Quine consiga explicar de forma convincente *cómo* es posible que aquélla pueda efectivamente desempeñar las tareas normativas clásicas. Las principales dificultades surgen cuando nos preguntamos por la justificación de las reglas metodológicas a observar por los científicos. La respuesta de Quine, acorde con una concepción puramente instrumental de la racionalidad, es que esas reglas se justifican por su validez como medios para el logro de los fines de la ciencia. Si suponemos, como Quine, que éstos vienen dados por la definición del *juego* de la ciencia³, entonces la justificación de las reglas se convierte en una

¹ O ese astuto pariente de este último, el convencionalismo, que sorprendentemente ha sabido ganarse mejor fama entre los filósofos analíticos.

² Cfr. Rodríguez Alcázar (1994).

³ Quine se refiere en ocasiones a la ciencia como uno de los *juegos de lenguaje* wittgensteinianos. Cfr. Quine (1990), 18 y 43.

cuestión empírica, naturalizable sin problemas: estaremos justificados para elegir aquellas reglas que, dada nuestra información actual, es más probable que contribuyan más eficazmente al logro de los objetivos propios de la ciencia.

En otros lugares he argumentado, contra esta “solución” quineana, que una epistemología genuinamente normativa no puede prescindir de la justificación de los objetivos científicos.⁴ En este punto se resume, a mi juicio, el principal reto para una epistemología naturalizada. También, como señalaré a continuación, para una filosofía naturalista de la ciencia.

2. El problema de la normatividad en la filosofía naturalista de la ciencia

Como ya he señalado, el naturalismo ha chocado tradicionalmente con algunos de los principios básicos de la tradición analítica, como la distinción analítico/sintético y la paralela diferenciación entre las tareas propias de los científicos y las de los filósofos. La situación no es diferente en el caso de la filosofía de la ciencia. En este terreno, el naturalismo ponía en cuestión uno de los pilares de aquella disciplina en su época dorada: la distinción entre los contextos de descubrimiento y justificación. Tal hecho no es sorprendente, pues una distinción tajante entre estos contextos sólo es posible si previamente se aceptan las nítidas diferenciaciones entre lo *a priori* y lo *a posteriori* y entre lógica y psicología tan caras a la filosofía analítica clásica.

La distinción entre contexto de descubrimiento y contexto de justificación pretendía diferenciar dos actitudes muy diferentes en la reflexión en torno a la ciencia: la descriptiva y la normativa. En virtud de la distinción, las relaciones entre las ciencias de la ciencia y la filosofía de la ciencia se reducen al mínimo: en tanto que aquéllas se ocupan de decir cómo es la ciencia desde diferentes perspectivas (histórica, sociológica, psicológica), a la filosofía de la ciencia le interesa más bien cómo debe ser⁵. Y, desde Hume y Moore, la tradición empirista ha dado por descontado que los conocimientos sobre el *es* no pueden decidir los debates sobre el *debe*. Entre los filósofos de la ciencia de la primera mitad del siglo predominó la convicción de que les era posible desarrollar su tarea mediante (poco más que) análisis lógico, recurriendo si acaso a ciencias como la historia para ilustrar algunas de las conclusiones previamente alcanzadas. El reto para los naturalistas consistía en mostrar que los resultados de (alguna de) las *ciencias de la ciencia* eran pertinentes para la filosofía de la ciencia. Eso es muy fácil si se altera el carácter de aquélla hasta convertirla en otra ciencia de la ciencia o conglomerado de ellas, pues entonces no hay ninguna brecha entre el *es* y el *debe* que sea necesario saltar⁶. Más difícil resulta convencer de que una

⁴ Cfr. Rodríguez Alcázar (1996); (en prensa), cap. 4.

⁵ Cfr. Reichenbach (1938), 5.

⁶ En esto desembocan finalmente, a mi juicio, los planteamientos de uno de los filósofos

filosofía de la ciencia que sustituye los resultados del análisis lógico por la indagación *a posteriori* de los procesos cognoscitivos individuales y colectivos puede seguir realizando las tareas normativas tradicionalmente encomendadas a la filosofía de la ciencia⁷. Creo que, como en el caso de la epistemología naturalizada, los problemas de la filosofía naturalista de la ciencia se ponen de manifiesto paradigmáticamente en la actitud de sus diversos representantes con respecto a los objetivos de la ciencia. La situación puede resumirse del modo siguiente. Los filósofos naturalistas pretenden proporcionar, frente al escepticismo y el relativismo, una justificación no fundamentista del conocimiento científico. Ahora bien, si para establecer en qué medida están justificadas las leyes y teorías recurrimos a las normas del método científico, y si éstas se justifican como medios para el logro de ciertos objetivos, ¿podemos afirmar que hemos conseguido una genuina justificación en alguno de los niveles mencionados sin una justificación de las elecciones de objetivos científicos?

Así pues, si es posible una genuina *filosofía* naturalista de la ciencia (esto es, una filosofía de la ciencia normativa pero no apriorista), los filósofos naturalistas deben convencernos de que están en condiciones de dar cuenta del debate racional sobre los fines de la ciencia, o bien de que dicho debate es innecesario para evitar el relativismo. A continuación paso a examinar diversas respuestas naturalistas a este reto.

3. Seis concepciones, seis, de los objetivos de la ciencia⁸

A la hora de caracterizar una concepción de los objetivos de la ciencia hay al menos cuatro preguntas pertinentes. La combinación de respuestas a cada una de esas preguntas produce diversas concepciones de dichos objetivos. Estas cuatro preguntas son:

- (1) ¿Cambian los objetivos de la ciencia con el tiempo?
- (2) ¿Hay un objetivo al que puedan reducirse o estén subordinados los demás?
- (3) ¿Hay criterios racionales para preferir unos objetivos a otros?
- (4) ¿Son los objetivos epistémicos los únicos objetivos de la ciencia?

Una respuesta negativa a (1) hará que califiquemos a quien la da como “esencialista”. Se puede ser esencialista con respecto a otros aspectos de la

naturalistas considerados en este trabajo, Ronald Giere.

⁷ Moulines (1995) pretende disolver este problema negándose a plantearlo en términos de la oposición descriptivo/normativo; a su juicio, la filosofía de la ciencia pertenecería a una tercera categoría, la de las disciplinas hermenéuticas. Critico la propuesta de Moulines en Rodríguez Alcázar (en prensa), cap. 1.

⁸ Utilizaré indistintamente las expresiones “metas”, “fines”, “objetivos” y “valores”, que consideraré extensionalmente equivalentes, aunque la última difiera intensionalmente de las anteriores. Así, la simplicidad ha sido considerada frecuentemente un “objetivo” de la ciencia, en tanto que muchos científicos la han perseguido, y también un “valor”, cuando se la ha utilizado como criterio de elección entre hipótesis.

ciencia, además de los valores. En particular, se puede ser esencialista con respecto al método⁹. Pero es posible ser esencialista sólo con respecto a los fines o valores de la ciencia; un esencialista de este tipo sostendría que aquéllos permanecen a lo largo del tiempo, pero habrían cambiado los juicios de las comunidades científicas acerca de las mejores estrategias para alcanzarlos.

Etiquetaré las dos respuestas posibles, positiva y negativa, a la pregunta (2), respectivamente como “monismo” y “pluralismo”. Es, naturalmente, posible que un esencialista sea tanto monista como pluralista. Pues alguien que sostenga que los objetivos de la ciencia no cambian puede creer bien en una pluralidad irreductible de objetivos permanentes, bien en un único objetivo principal a lo largo del tiempo¹⁰.

La formulación de (2) incluye, en realidad, dos preguntas. La primera es si se puede hablar de un único objetivo para la ciencia, o uno al que los demás puedan reducirse. Llamaré “monismo *fuerte*” a la respuesta positiva a esa pregunta. La segunda pregunta es si, aún aceptando la existencia de una pluralidad de valores no reductibles, cabe hablar de una *jerarquía* estática entre ellos, de forma que pueda identificarse un objetivo al que siempre están subordinados los demás. Me referiré a esta posición como “monismo *débil*.” La respuesta que se dé a (3) depende, como ya sabemos, de que se acepte o no una concepción puramente instrumental de la racionalidad. Finalmente, las respuestas a (4) dividen a sus proponentes respectivos en “internalistas” y “externalistas”.

Recordemos: el filósofo naturalista de la ciencia, que quiere salvaguardar el carácter normativo de su disciplina, pretende evitar el relativismo sin recaer en el fundamentismo, siendo al mismo tiempo consecuente con la información de las distintas ciencias sobre la práctica y el conocimiento científicos. Pues bien: algunas combinaciones de respuestas a las cuatro preguntas anteriores no cumplen esas condiciones. Así, una combinación atractiva para sociólogos e historiadores constructivistas por su compatibilidad con los resultados de sus investigaciones (o, al menos, con la lectura que ellos hacen de esos resultados), sería la siguiente: antiesencialismo, pluralismo, concepción instrumental de la racionalidad y externalismo. Pero esta combinación tiene el inconveniente, para el filósofo de la ciencia, de que conduce directamente al relativismo. En efecto, si las axiologías cambian de unas comunidades científicas y de unas épocas a otras; si, incluso entre los miembros de una misma comunidad científica, no está fijada una jerarquía entre los criterios de valoración que permita establecer

⁹ Un ejemplo es Worrall (1988); también, como veremos, Kitcher vincula su esencialismo con respecto a las metas con un esencialismo del método.

¹⁰ Kuhn (1977) es de la primera opinión; la segunda coincide, como se pondrá de manifiesto más adelante, con el punto de vista de Kitcher.

automáticamente qué hacer en caso de conflicto entre esos criterios; si admitimos que entre los criterios pertinentes se encuentran valores no epistémicos y si, finalmente, aceptamos que no existe, frente a toda esa pluralidad de criterios de valoración, ningún metacriterio que permita juzgarlos y establecer, en cada contexto, un orden de prelación entre ellos, entonces el resultado previsible será la inconmensurabilidad radical de teorías y métodos. Pues siempre será posible para el defensor de una teoría o de una metodología encontrar algún criterio, epistémico o práctico, con respecto al cual su teoría o su método sea superior a los rivales, y no habrá, en el contexto de una concepción meramente instrumental de la racionalidad, recursos para criticar su elección.

¿Tendremos que resignarnos a esta situación? ¿Conducen a este relativismo irremediamente los hallazgos de la historia y la sociología de la ciencia? Si bien los representantes del constructivismo social saludan el relativismo como un requisito metodológico imprescindible,¹¹ entre los filósofos el relativismo suele parecer tan poco deseable que habitualmente apuran los intentos de evitarlo. Quizás sea posible encontrar una combinación de respuestas a las preguntas anteriores que evite desembocar en el relativismo. ¿Pero existe esa combinación?

Una posibilidad extrema (que calificaré de *ultra-racionalista*) consiste en responder a todas las preguntas anteriores lo contrario que los relativistas. De ser esa combinación viable, nos libraría, ciertamente, del relativismo. Pero esa opción *ultra-racionalista* parece poco creíble en su integridad. En efecto, su defensor tendría un arduo trabajo contestando a la documentación histórica de los (aparentes) cambios en los fines de la ciencia a lo largo del tiempo, a la documentación sociológica de la pluralidad de intereses dentro de cada comunidad y al externalismo tan extendido entre historiadores y sociólogos. Naturalmente, no puedo acompañar al *ultra-racionalista* a todas sus batallas para comprobar cómo le va en cada una (el debate histórico y sociológico promete mantenerse vivo todavía algunos años y el plazo de que dispongo para entregar este trabajo es considerablemente menor). Sí podemos preguntarnos ya, sin más espera, si realmente resulta imprescindible refutar la primera combinación en cada uno de sus puntos para evitar el relativismo. Tres filósofos naturalistas, Giere, Kitcher y Laudan están convencidos de que no es necesario, construyendo así tres propuestas intermedias entre la claramente relativista y su contraria, la *ultra-racionalista*. Las tres difieren entre sí, así que es interesante comprobar en cada caso si se realiza el objetivo de evitar el relativismo y, de ser así, a qué coste.

La propuesta de Ronald Giere comparte bastantes puntos con la relativista. Giere acepta la concepción instrumental de la racionalidad, que

¹¹ Cfr. Barnes y Bloor (1982).

considera un rasgo definitorio del naturalismo¹². También acepta la influencia de una gran diversidad de valores e intereses sobre el desarrollo de la ciencia, lo cual nos inclinaría a tenerlo por un pluralista si no fuera porque, como voy a señalar enseguida, todos esos valores resultan estar subordinados a un valor epistémico predominante, lo que me obliga a etiquetarlo como un monista débil. Por último, admite que entre esos valores e intereses algunos no son epistémicos, lo que nos autoriza a etiquetarlo como externalista. Es más, Giere considera que a los efectos de una teoría cognoscitiva, explicativa, de la ciencia (a lo que él reduce la filosofía de la ciencia), ni siquiera sería necesario distinguir entre valores epistémicos y valores o intereses sociales. Dado que, según Giere, las personas no distinguen claramente entre los diversos factores que influyen en su juicio, sólo es posible explicar adecuadamente la conducta de los científicos utilizando un modelo de toma de decisiones que tenga en cuenta ambos tipos de valores¹³.

Pese a estas declaraciones externalistas, Giere se desmarca del relativismo de los constructivistas sociales¹⁴. Ahora bien, ¿es posible evitar el relativismo a partir de esa combinación de externalismo, racionalidad instrumental y aparente pluralismo que acabo de presentar? ¿Cuál es el elemento en su lista de respuestas que le permitiría hacerlo? Se trata de un ingrediente esencialista (que, de paso, sirve para alterar drásticamente el carácter inicialmente pluralista de la propuesta). Para dar sentido a la idea de *progreso* científico, Giere introduce un valor epistémico privilegiado, permanente, que se impone sobre los intereses de diverso tipo que él reconoce en los científicos. Este criterio es la *semejanza* entre los modelos científicos y los sistemas reales. Si bien la semejanza ha sido frecuentemente objetada por los filósofos, que la han considerado demasiado vaga, Giere considera suficiente el respaldo que le proporcionan las ciencias cognoscitivas.

El peso de la semejanza es tal que los demás valores e intereses de los científicos sólo entran en juego como complementarios de aquel criterio, cuando la semejanza por sí sola no determina la elección entre modelos rivales¹⁵. En aquellos casos en que un modelo es claramente más semejante al sistema real que el alternativo, el criterio de la semejanza eclipsa completamente a los demás¹⁶. Esta solución es claramente esencialista, pues el peso del criterio vendría dado por la naturaleza representacional de la ciencia, algo no sujeto a cambio con el paso del tiempo. A la vez, matiza de forma

¹² Giere (1988), 17, 194. Consecuentemente, Giere, de forma semejante a Quine, sólo admite un sentido legítimo en el que la filosofía de la ciencia puede considerarse normativa: en tanto que puede ayudarnos a decidir cuáles son los mejores *medios* para alcanzar las metas de los científicos.

¹³ Giere (1988), 196.

¹⁴ Cfr. Giere (1988), 24, 31, 78.

¹⁵ Cfr. Giere (1988), 195.

¹⁶ Cfr. Giere (1988), 263.

importante el pluralismo de Giere, pues, tras haberse negado éste (por su adhesión a la concepción puramente instrumental de la racionalidad) a establecer criterios para comparar el enjambre de intereses y valores, epistémicos o no, que influyen en los científicos, ahora los eclipsa a todos introduciendo un valor predominante.

La solución de Giere presenta, al menos, dos problemas importantes. Uno es el alto coste de su esencialismo. Pues, como comprobaremos enseguida al ocuparnos de la polémica entre Kitcher y Laudan, el esencialismo se enfrenta con argumentos históricos dignos de consideración y, además, resulta poco consistente con el talante naturalista. Pero, dejando estas objeciones aparcadas por el momento, hay un segundo problema más importante: a pesar de ese costoso recurso al esencialismo, Giere no consigue evitar el relativismo. En efecto, los filósofos se han quejado a menudo, con razón, de que la semejanza es un criterio vago: todo es semejante a todo en algún respecto y el juicio acerca del parecido entre la representación y lo representado depende de convenciones culturales¹⁷. Así (por utilizar el ejemplo favorito de Giere cuando explica su noción de modelo científico) no tiene sentido preguntarse en términos absolutos qué mapa se parece más a una porción del mundo. Por tanto, aunque admitamos, con Giere, que los científicos utilizan la semejanza como criterio de elección de modelos, también hemos de reconocer que esos científicos asociarán con el criterio, en cada caso, características diferentes, por lo que es muy posible que discrepen con respecto a qué modelo guarda una mayor semejanza con el sistema real de que se trate. Así pues, Giere no puede garantizar que no ocurra precisamente lo anunciado por los sociólogos del conocimiento científico: que los científicos atribuyan mayor o menor semejanza con la realidad a unos u otros modelos precisamente en función de sus propios intereses (unos intereses que Giere se niega a juzgar, dada su adhesión a una concepción puramente instrumental de la racionalidad). En resumen, a Giere no le basta el recurso esencialista a la semejanza para evitar el relativismo.

¿Dónde está el error que conduce a Giere hasta el relativismo? Alguien que comparta su querencia por el esencialismo puede argumentar que el problema está en la elección del objetivo permanente de la ciencia. Tal es la opinión de Philip Kitcher, aunque éste, además, argumenta en favor de su propia propuesta axiológica con recursos que lo alejan de la concepción meramente instrumental de la racionalidad.

En polémica con Laudan, Kitcher señala la necesidad de una axiología permanente¹⁸. Esa necesidad se justificaría porque los cambios de objetivos a lo largo de la historia pondrían en peligro la racionalidad del cambio científico, que tanto Laudan como Kitcher pretenden preservar. De acuerdo

¹⁷ Cfr. Rosenberg (1996), 12; Barnes (1977), cap. 1.

¹⁸ Cfr. Kitcher (1993), 157 ss.

con el “modelo reticulado” de Laudan, cada estadio de la evolución de una disciplina científica puede analizarse en tres niveles: el de las teorías, el de los métodos y el de los fines o valores (“fáctico”, “metodológico” y “axiológico”, en su terminología). Las transiciones de un estadio a otro serían racionales porque en ellas no se produce un cambio simultáneo en los tres niveles; antes bien, la modificación de cualquiera de ellos presupone siempre el consenso de los científicos acerca de los otros dos. Pero, objeta Kitcher, si ha de preservarse la racionalidad del proceso, es necesario que permanezcan a lo largo de él unos principios metodológicos estables que lo gobiernen. Éstos deberían conformar una metodología permanente, integrada por las estrategias para el logro racional de un objetivo científico fijo, no sujeto a ese reajuste constante entre teorías, métodos y fines descrito por el modelo reticular de Laudan. Ese objetivo es, según Kitcher, el logro de verdades relevantes (“significant”)¹⁹. Aquí estaría el elemento esencialista y monista de la propuesta de Kitcher. El monismo, sin embargo, se relaja un tanto si tenemos en cuenta que lo que se considera “relevante” puede variar según el contexto y según la época. Así, Kitcher concordaría con Laudan en que se producen modificaciones en la axiología incluida en cada uno de los sucesivos estadios del proceso científico, pero esas axiologías no serían más que conjuntos de sub-metas que proporcionarían concreción al objetivo permanente de la búsqueda de verdades relevantes.

Naturalmente, la propuesta (a la vez descriptiva y normativa) de la verdad relevante como objetivo de la ciencia enfrenta a Kitcher con antirrealistas como van Fraassen y el propio Laudan²⁰, pero lo que en este momento nos concierne no es la legitimidad de un determinado objetivo, sino la necesidad o no de suponer que la ciencia está regida por una axiología monista y esencialista. ¿Tiene razón Kitcher al sostener que la racionalidad del cambio científico está en peligro en ausencia de una axiología estable y débilmente monista?

A mi juicio, no la tiene. De hecho, la posición de Kitcher en este punto resulta bastante sorprendente en un declarado naturalista. Pues la bien fundada desconfianza de los naturalistas hacia toda forma de conocimiento *a priori* y de discurso trascendental parece difícilmente conciliable con la postulación de unas reglas y unas metas que escaparían a toda posible revisión y duda. Si los naturalistas pueden convivir con la falibilidad y la revisabilidad de cualquier porción de conocimiento científico, también deben convivir con la falibilidad y la revisabilidad no sólo de sus concepciones sobre las reglas y los objetivos de la ciencia, sino también de las reglas y los objetivos mismos, en tanto que éstos son resultado del quehacer humano. El examen histórico ofrece fuertes indicios a favor de la existencia de importantes cambios en los métodos y los

¹⁹ Cfr. Kitcher (1992), 102; (1993), 94-95.

²⁰ Cfr., p. ej., Kitcher (1993), 140 ss.

finde de la ciencia. Pero aunque fuera posible encontrar, como Kitcher sostiene, algunos elementos invariantes, ello no debería considerarse equivalente al hallazgo de una esencia (siquiera mínima) de la actividad científica, sino a la mera constatación *a posteriori* de que, hasta el momento, ciertos ingredientes han permanecido. Los filósofos naturalistas deberían estar en condiciones de explicar a los relativistas cómo es posible salvar la racionalidad de las transiciones científicas a pesar de que, como éstos sostienen, quizás se hayan producido cambios en cada uno de los niveles de la ciencia y, lo que es más importante, a pesar de que, se hayan producido o no, cada ingrediente de todos esos niveles está y debe estar, en principio, abierto a revisión y escrutinio. Los científicos y los filósofos “naturalizados” se sitúan *in media res* en un entramado de creencias fácticas, metodológicas y axiológicas que aceptan en líneas generales y que proporcionan el punto de partida de cualquier debate en cualquiera de esos niveles. La racionalidad de las transiciones queda garantizada en cada caso por la continuidad de la gran mayoría de los ingredientes del edificio científico. No sólo porque, como resultaría de un despliegue genérico del modelo de Laudan, los niveles fáctico y axiológico, por ejemplo, permanecen estables mientras se revisa el metodológico; también porque la mayoría de los principios metodológicos siguen asimismo incuestionados mientras se produce el debate en torno a alguno de ellos. Esas normas que permanecen en cada fase fuera de discusión (y que pueden gozar, según los casos, de una larga vida, incluso haber acompañado a la ciencia desde sus inicios) garantizan la continuidad en las transiciones de la forma requerida por Kitcher. Pero éste, en tanto que naturalista, no está legitimado para extraer algunas de ellas de la inestable trama de las creencias científicas y declararlas inmunes a la crítica. Ni siquiera le autorizaría a hacer tal cosa un inequívoco dictamen de la historia en favor de la permanencia *de facto* de algunas normas y objetivos de la ciencia (un dictamen inequívoco que, en cualquier caso, no está, ni mucho menos, disponible).

Ciertamente, Kitcher tiene más éxito que Giere a la hora de evitar el relativismo. Pero la clave de su éxito no está ni en su esencialismo ni en su monismo débil. De hecho, éstos solos no bastarían (como no le bastan a Giere), pues los relativistas siempre podrían objetar que, si lo que se consideran verdades *relevantes* cambia con la época y el contexto, el recurso de Kitcher al realismo no basta para restringir en la medida en que aquél quisiera los contenidos admisibles de la ciencia. Así pues, cuando el monismo tanto de Giere como de Kitcher se matizan para hacer justicia a la realidad histórica y sociológica de la ciencia, parecen conducir directamente al relativismo (un relativismo que asoma en cuanto nos preguntamos qué entendemos por “semejanza” entre dos modelos o qué verdades son “relevantes”). Lo que salva a Kitcher del relativismo es su adhesión a una concepción de la racionalidad no exclusivamente instrumental, que le permite dar sentido a la

posibilidad de debatir racionalmente también aquellas sub-metas a la luz de las cuales se decide qué verdades son “relevantes” y cuáles no²¹. Pero esta vacuna frente al relativismo es igualmente eficaz en ausencia del compromiso adicional de Kitcher con el esencialismo y el monismo débil. Ese compromiso, además de intrínsecamente problemático, resulta, pues, innecesario para la búsqueda naturalista de una respuesta no fundamentista al relativismo.

La propuesta de Laudan proporciona los ingredientes necesarios para hacer frente al relativismo sin los costes del esencialismo y del monismo. Si bien Laudan (1977) adoptó inicialmente un punto de vista esencialista y monista, caracterizando el progreso científico en términos de la eficacia en la resolución de problemas, posteriormente hace depender la posibilidad del progreso y del consenso científicos de una concepción de la racionalidad no puramente instrumental que permite el escrutinio de los objetivos de la ciencia: cfr. Laudan (1984). El modelo de Laudan hace posible el debate racional acerca de cada uno de los niveles científicos, dados un cierto momento histórico y un cierto contexto. Dado que las diferencias en información fáctica y en recursos metodológicos pueden hacer necesario modificar los objetivos según el momento y el contexto, Laudan, para desesperación de sus críticos esencialistas, no puede consistentemente señalar sin más cuál es o cuáles son los objetivos de la ciencia. Éstos cambian; pero su posición no es relativista, pues permite la posibilidad de decidir racionalmente cuáles son los objetivos más adecuados en cada contexto y momento²².

Ahora bien, la propuesta de Laudan incluye un elemento problemático desde el punto de vista de algunos resultados de la historia y la sociología de la ciencia, así como completamente superfluo para el propósito de evitar el relativismo. A diferencia de Kitcher y Giere, Laudan adopta una posición internalista con respecto a los objetivos de la ciencia²³. En realidad, Laudan no

²¹ Cfr. Kitcher (1993), 179, n. 4.

²² Aunque Laudan (1984) no se pronuncia por ninguna axiología particular, sí lo hace negativamente, en consonancia con su antirrealismo, para rechazar la verdad como valor epistémico. A mi juicio, ésta es la parte menos afortunada del libro, pues no consigue mostrar que el rechazo del realismo se sigue de las consideraciones previas. Mi propio punto de vista es que no hay razones sólidas para excluir a la verdad como *un* valor científico, aunque sí para pensar que no es el único valor epistémico importante y que, en ciertos contextos, ocupa una posición subordinada con respecto a otros.

²³ Es necesario, con todo, referirse con precaución al “externalismo” de Kitcher. Por un lado, en el plano descriptivo, éste reconoce que los objetivos perseguidos por los científicos no se reducen a los epistémicos; sostiene, además, que el rechazo naturalista del *a priori* obliga a reconocer la dimensión social del conocimiento; finalmente, se interesa por cómo deberían diseñarse las instituciones sociales para favorecer el logro de los objetivos epistémicos de la ciencia: cfr. Kitcher (1992), 82, 110-111; (1993), 61, 72-73, 92. Sin embargo, su propuesta normativa de objetivos para la ciencia se reduce a la defensa de un valor epistémico, la producción de verdades relevantes, aplazando la discusión de los valores no epistémicos *ad calendas graecas*: cfr. Kitcher (1992), 107; (1993), 92.

discute con historiadores y sociólogos acerca del grado de influencia de factores y, en particular, valores “externos” sobre el proceso científico. Más bien adopta una actitud “lakatosiana”: puesto que su intención es dar cuenta de la racionalidad, el progreso y el consenso científicos, y dando por supuesto que éstos se ponen en peligro con la aparición en escena de valores no epistémicos, Laudan simplemente los excluye de su modelo reticulado²⁴.

El temor a que el reconocimiento de la influencia de valores no epistémicos conduzca al relativismo tiene cierto fundamento. Recordemos que tanto la combinación relativista de respuestas a las cuatro preguntas de marras como la de Giere (que acabé acusando también de relativista) incluyen el externalismo entre sus ingredientes. Téngase en cuenta, además, que los sociólogos constructivistas hacen del externalismo y del relativismo, simultáneamente, dos de sus principales lemas. Pero el externalismo sólo pone en peligro nociones normativas como la de progreso cuando se combina, como ocurre en los casos mencionados, con una concepción puramente instrumental de la racionalidad. Laudan no debería temer que su modelo reticulado acabe permitiendo cualquier dirección en el desarrollo científico sólo porque incluya valores no epistémicos, a condición de que tengamos fundadas esperanzas de extender el debate racional hasta éstos²⁵. Esa extensión convertiría seguramente al modelo reticulado en un modelo más adecuado desde el punto de vista histórico y sociológico, aunque no es mi propósito aquí terciar en la compleja (y abierta) disputa entre internalistas y externalistas. En cualquier caso, permitiría que la supervivencia futura de la filosofía de la ciencia como disciplina normativa no dependa de los descubrimientos de historiadores y sociólogos, pues mostrar que la elección de los contenidos, los métodos o los objetivos epistémicos de la ciencia está determinada en algún grado y en algunos contextos por la elección de valores no epistémicos no tiene por qué convertir a los primeros en inconmensurables, a condición de que no pensemos que los segundos lo son siempre. En cambio, los filósofos de la ciencia habrían de acostumbrarse a una mayor dependencia de los resultados de la teoría ética y otras ramas de la filosofía práctica.

De esta forma llegamos a la combinación de respuestas que hoy por hoy me parece más plausible. Esa combinación aparece recogida, como una propuesta naturalista para el futuro, junto con las anteriormente discutidas, en la siguiente tabla:

²⁴ Cfr. Laudan (1984), 138-139.

²⁵ Shrader-Frechette (1989) propone incluir valores morales en el modelo de Laudan. A pesar del interés de su propuesta, he señalado algunas insuficiencias de la misma: cfr. Rodríguez Alcázar (en prensa), cap. 6.

	ESENCIALISMO	MONISMO	RAC.SUSTANT.	INTERNALISMO
RELATIVISTA	NO	NO	NO	NO
ULTRA-RAC.	SÍ	FUERTE	SÍ	SÍ
GIERE	SÍ	DÉBIL	NO	NO
KITCHER	SÍ	DÉBIL	SÍ	NO (...)
LAUDAN	NO	NO	SÍ	SI
NAT. FUTURO	NO	NO	SÍ	NO

La combinación que apoyo comparte con la relativista el antiesencialismo, el pluralismo y el externalismo, pero el relativismo se evita gracias al rechazo de la racionalidad instrumental y el concurso de la teoría ética. Aunque no puedo detenerme en este asunto, resulta aleccionador que algunos de los representantes más destacados de aquella en los últimos años se hayan alejado del no-cognoscitismo más o menos emotivista compartido, curiosamente, por filósofos positivistas y sociólogos relativistas con respecto a los valores morales. Estas consideraciones nos conducen a otra de las características de una filosofía naturalista de la ciencia viable en el futuro: su atención a los resultados de la filosofía práctica. Los epistemólogos naturalistas, que ya escandalizaron el vecindario analítico al resucitar el mismo psicologismo que Frege dejara bien enterrado y al burlarse de algunas distinciones que se consideraban nítidamente trazadas (analítico-sintético, descubrimiento-justificación), habrán de ensuciar aún más su hoja de servicios uniéndose a aquellos filósofos morales que andan profanando otro de los templos del análisis: el consagrado a la separación del *es* y el *debe* y a la alabada derrota por G.E. Moore de una presunta falacia.

BIBLIOGRAFÍA

- Barnes, Barry (1977): *Interests and the Growth of Knowledge*. Londres: Routledge and Kegan Paul.
- Barnes, B., y Bloor, D. (1982): "Relativismo, racionalismo y sociología del conocimiento", en González García, M.I., López Cerezo, J.A. y Luján, J.A. (eds.) *Ciencia, tecnología y sociedad. Lecturas seleccionadas*. Barcelona, Ariel, 1997, 27-48.
- Giere, R. (1988): *La explicación de la ciencia*. Trad. cast. de Claudia Elisa Gidi Blanchet. Méjico: Consejo Nacional de Ciencia y Tecnología, 1992.
- Kitcher, P. (1992): "The Naturalists Return", *The Philosophical Review*, vol. 101, n. 1, 53-114.
- (1993): *The Advancement of Science*, Oxford: Oxford University Press.
- Kuhn, T.S. (1977): "Objetividad, juicios de valor y elección de teoría", en *La tensión esencial*. Méjico: FCE, 1982.
- Laudan, L. (1977): *El progreso y sus problemas*. Madrid: Encuentro, 1986.
- (1984): *Science and Values: The Aims of Science and Their Role in Scientific Debate*. Berkeley: University of California Press.
- (1990): "Normative Naturalism", *Philosophy of Science*, 57, 44-59.
- Moulines, C.Ulises (1995): "La filosofía de la ciencia como disciplina hermenéutica". *Isegoría*, 12, 110-118.
- Quine, W.V. (1990): *La búsqueda de la verdad*. Trad. cast. de F. Javier Rodríguez Alcázar. Barcelona: Crítica, 1992.

- Reichenbach, Hans (1938): *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: the University of Chicago Press.
- Rodríguez Alcázar, F.J. (1994): "La polémica sobre epistemología naturalizada y normatividad". *Agora*, 13/1, 95-119.
- (1996): "Naturalism Radicalized", *Metaphilosophy*, vol. 27, n. 4, 356-380.
- (en prensa): *Ciencia, valores y relativismo: una defensa de la filosofía de la ciencia*.
- Rosenberg, A. (1996): "A Field Guide to Recent Species of Naturalism", *British Journal for the Philosophy of Science*, 47, 1-29.
- Shrader-Frechette, Kristin (1989): "Scientific Progress and Models of Justification: A Case in Hydrogeology", en Goldman, S.L. (ed.) *Science, Technology and Social Progress*. Bethelhem, Lehigh University Press.
- Worral, John (1988): "The Value of a Fixed Methodology", *British Journal for the Philosophy of Science*, 39, 263-275.

F. Javier Rodríguez Alcázar
Universidad de Granada
e-mail: frodiguez@goliat.ugr.es



INCONMENSURABILIDAD LÉXICA Y ESPECIALIZACIÓN EN EL ÚLTIMO KUHN

PATRICIA RODRÍGUEZ FERNÁNDEZ
Universidad de Santiago de Compostela

ABSTRACT: The aim of this paper is to tackle Kuhn's treatment of the notion of incommensurability at the last stage of his epistemology, in particular the link of this notion with the specialization. Though Kuhn does not abandon the diachronical projection of the concept of "incommensurability", in his last articles he deviates attention toward the application of such notion to the analysis of the relation existing between the different specialties emerged from a common trunk.

KEYWORDS: Epistemology, incommensurability, specialization, local holism, lexicon.

1. Introducción

Mi propósito aquí es recomponer, a partir de una pequeña selección de artículos pertenecientes a la etapa del pensamiento de Kuhn comprendida entre los años 1982 y 1986, las distintas piezas que configuran la nueva imagen de la inconmensurabilidad vinculada al fenómeno de la especialización del conocimiento. En el contexto de la obra de Kuhn, la noción de "inconmensurabilidad" vertebró toda su concepción epistemológica, si bien su caracterización se ha visto sometida a cambios importantes a lo largo de su obra. En textos como "El camino desde «La Estructura»" y "Afterwords" (ver Bibliografía) podemos constatar un cierto cambio de rumbo en la caracterización de las revoluciones científicas: ya no son interpretadas como rupturas en la línea de desarrollo de la ciencia, sino que se identifican con procesos de especialización del conocimiento científico. Para ilustrar este tipo de cambio recurre constantemente a ejemplos tomados de la biología evolutiva: "El equivalente biológico al cambio revolucionario no es la mutación, como pensé durante muchos años, sino la especiación (KUHN, 1994: p. 37)." No obstante, los ejemplos mencionados por Kuhn de procesos de especialización dentro del conocimiento científico son escasos, de ahí que en ocasiones surjan dificultades en su comprensión. He aquí una de ellas:

Kuhn no deja claro si su intención fue utilizar la noción para referirse al proceso mediante el cual se constituyeron las distintas disciplinas en que se ha ido fragmentando el saber occidental, o si en realidad se refería sólo a la proliferación de múltiples áreas dentro de cada una de las disciplinas que componen la ciencia. De la lectura de sus textos parece inferirse que el alcance del concepto de especialización es total, sobre todo cuando se refiere a la filosofía como el tronco a partir del cual se han ido desgajando las ciencias. No obstante, los ejemplos de cambio taxonómico local utilizados en los textos en que trata de caracterizar este concepto, se circunscriben a una única disciplina (la introducción de la segunda ley del movimiento de Newton, sobre la base de un vocabulario antecedente previamente disponible en el ámbito de la física; descubrimiento de los tres estados de agregación del agua, en el ámbito de la química, etc.). Trataré de defender que la relevancia del fenómeno de especialización es mucho mayor cuando se extiende a la totalidad del conocimiento humano, y no sólo a las distintas áreas que surgen en el seno de cada una de las disciplinas científicas.

Me ha parecido interesante comparar el concepto de especialización de Kuhn con el propio de la Concepción Estructuralista, no con la intención de corregir una confusión inexistente entre ambos, sino con el objeto de mostrar que ambos usos son útiles cuando intentamos ofrecer una imagen completa de la ciencia. El concepto de Kuhn nos permite interpretar el alto grado de fragmentación del saber científico en la actualidad, como una consecuencia de la aspiración del ser humano a ensanchar los dominios del saber, extendiendo su poder hasta los límites de sus capacidades. La progresiva división del trabajo científico obedece a un intento de llegar cada vez más cerca de los fenómenos. Por su parte, el concepto estructuralista daría cuenta de la ramificación sucesiva de un núcleo teórico, en su intento de acercarse cada vez más a la realidad que pretende explicar. En el fondo, ¿no se trata del mismo mecanismo, actuando sobre ámbitos distintos? Ambos conceptos de especialización pueden ilustrarse mediante el uso de la metáfora del árbol invertido, cuyas ramas simbolizarían el permanente esfuerzo del ser humano por lograr que sus teorías lleguen alguna vez a tocar la realidad. Hemos tenido que pulir constantemente nuestras herramientas conceptuales y lingüísticas, para aumentar su eficacia. De ahí que, en beneficio de la precisión, haya sido necesario introducir vocabulario especializado para dar cuenta de nuevos ámbitos de fenómenos o para acceder a un nuevo nivel de investigación de la realidad.

2. Vinculación de la inconmensurabilidad con la especialización en Kuhn

En el artículo de 1990 "El camino desde «La estructura»", Kuhn distingue dos aplicaciones posibles de la noción de "inconmensurabilidad" (ver p. 35 y sgs.). La primera ha sido ampliamente desarrollada a lo largo de toda su obra y se circunscribe a los análisis diacrónicos de la ciencia. Se habla

de inconmensurabilidad entre teorías rivales que han protagonizado etapas sucesivas de ciencia normal de una determinada disciplina: la astronomía ptolemaica y la copernicana, la teoría aristotélica del movimiento y la teoría del ímpetus, esta última y la física newtoniana, la teoría del flogisto y la química moderna... En su segunda aplicación, el concepto permitiría dar cuenta de la relación existente entre dos o más especialidades científicas localizadas en un determinado momento histórico, pongamos por caso: la Física de Partículas Elementales, la Física Clásica Macroscópica y la Física Relativista. Si alguien se propone reconstruir, mediante un análisis sincrónico, la imagen que ofrece la ciencia en nuestra época, o en una etapa reciente de su evolución, comprobará el alto grado de fragmentación a que se ha visto sometido el conocimiento científico. La división del trabajo cognitivo y la puesta en funcionamiento de todas las capacidades con que está dotado el hombre, con el fin de ensanchar los límites de nuestro conocimiento en todas sus dimensiones, ha tenido una contrapartida negativa: la ruptura de la comunicación entre las comunidades de especialistas que componen la población científica.

La diferencia fundamental entre las dos proyecciones del concepto de inconmensurabilidad no se halla, como pretende hacernos creer Kuhn, en la perspectiva –sincrónica o diacrónica– desde la cual se realiza el análisis. La diferencia reside, más bien, en el marco que se toma como referencia: la primera pretende dar cuenta de la relación existente entre las distintas teorías que se van sucediendo a lo largo de la historia de una disciplina; en cambio, la segunda pretende explorar las relaciones existentes entre diferentes disciplinas científicas o campos de conocimiento. La perspectiva diacrónica está también presente en esta segunda aplicación, ya que a través de ella Kuhn intenta ofrecer una imagen global de la ciencia, describiendo el proceso mediante el cual se constituyen las distintas especialidades que la componen:

Estas roturas de comunicación ocurren, por supuesto: son una característica significativa de los episodios referidos en *La estructura* como “crisis”. Yo los considero como síntomas cruciales del proceso, similar a la especiación, a través del cual emergen nuevas disciplinas, cada una con su propio léxico, y cada una con su propia área de conocimiento (Ibíd.: p. 40).

En cualquiera de las dos aplicaciones se maneja el concepto de inconmensurabilidad derivado de la formulación que el propio Kuhn designa como “inconmensurabilidad local”. Cada vez que tiene lugar una revolución científica, se lleva a cabo una reconstrucción de parte del vocabulario de la ciencia: se introducen nuevos grupos de términos interdefinidos, cuyos significados se estipulan conjuntamente. Estos grupos de términos no sólo no son traducibles al lenguaje previamente disponible; si los adoptamos, nos vemos obligados a abandonar aquella parte de nuestro viejo vocabulario que

se solapa con el nuevo en el mismo dominio. La inconmensurabilidad se produce, pues, como consecuencia de la necesidad de llevar a acabo ajustes locales del léxico, con el objeto de responder a las demandas propias de cada etapa de la historia de la ciencia.

En los últimos artículos de Kuhn es frecuente la caracterización de una teoría científica como un sistema léxico en el que se estructuran y clasifican los objetos de un determinado dominio. Las clases que integran cada sistema léxico condensan el significado de la realidad sobre la que se proyectan. El léxico de la ciencia se ha visto sometido a múltiples alteraciones a lo largo de su historia, como resultado de cambios drásticos que conllevan reestructuraciones profundas en alguna región local del vocabulario. La distinción que se establecía en la *Estructura de las Revoluciones Científicas* entre desarrollo revolucionario y desarrollo normal es equivalente a la que se hace ahora entre desarrollos con cambio taxonómico local y aquellos que no lo requieren.

En el contexto de la comparación entre disciplinas, la inconmensurabilidad es una consecuencia de la especialización, mecanismo aislante que contribuye a aumentar la fuerza cognitiva de la ciencia, permitiéndole tener el control sobre una extensión cada vez más amplia de fenómenos. Para explicar el modo en que se han ido constituyendo las diversas especialidades que integran la ciencia en la actualidad, Kuhn utiliza la metáfora del desarrollo biológico. En primer lugar, considera equiparables los procesos de especialización científica y de especiación biológica. Con frecuencia comprobamos que tras una revolución científica ha aumentado el número de campos de conocimiento, bien sea porque un nuevo vástago se ha desgajado del tronco paterno, o bien como resultado del nacimiento de una nueva especialidad en un área de solapamiento entre dos disciplinas previamente existentes:

Después de una revolución, hay usualmente (...) más especialidades cognitivas o campos de conocimiento de los que había antes. Bien una nueva rama se ha desgajado del tronco paterno, como lo han hecho repetidamente en el pasado especialidades científicas que se han separado de la filosofía y de la medicina; o bien una nueva especialidad ha nacido en un área de solapamiento aparente entre dos especialidades preexistentes, como ha ocurrido, por ejemplo, en el caso de la química física y de la biología molecular (Kuhn, 1994: p. 36).

En segundo lugar, este paralelismo entre el desarrollo científico y la evolución biológica se extiende también a las comunidades resultantes de un proceso de especialización. Quienes integran una comunidad de especialistas se aíslan de los restantes grupos de expertos mediante el uso de un léxico característico, que es esencial para la comprensión y evaluación de los

conocimientos obtenidos por el grupo y que, a su vez, obstaculiza la plena comunicación con los no iniciados. La contrapartida biológica es una población con un fondo genético común, el cual al tiempo que garantiza la autoperpetuación de la especie, hace posible su aislamiento respecto de grupos diferentes. Los campos cognitivos que conforman el mapa de la ciencia difieren en el léxico, debiendo buscarse el origen de tal diversidad en los procesos revolucionarios que han obligado a una reconstrucción de parte de nuestras taxonomías. Los responsables de estas innovaciones se separan de la comunidad científica de que formaban parte, aislándose como clase profesional e inaugurando una nueva especialidad dentro de la ciencia.

Intentemos describir, a grandes rasgos, el proceso de especialización científica, tal como lo concibe Kuhn. 1) En cada una de sus etapas de desarrollo, la ciencia cuenta con un vocabulario específico, que proviene de las teorías que se han propuesto como candidatas para la explicación de los fenómenos y que, hasta el momento, han demostrado tener éxito. 2) El período de crisis comienza cuando los fenómenos observados violan nuestras expectativas, produciéndose un desajuste entre nuestro aparato lingüístico y conceptual, y la realidad observada. 3) En circunstancias como éstas, los científicos se ven en la necesidad de buscar una explicación satisfactoria de los fenómenos anómalos, para lo cual no bastará con un enunciado aislado, sino que será necesario proponer alguna teoría alternativa. 4) La aplicación de los criterios de evaluación de la ciencia –exactitud, precisión, alcance, simplicidad, utilidad, consistencia, etc.– permite que la crisis se resuelva con el surgimiento de una nueva especialidad. La nueva teoría contará con un vocabulario antecedente, disponible a partir de teorías preexistentes, y un grupo de términos específicos, cuyo significado depende de ciertas leyes establecidas por ella misma. En este punto, Kuhn se hace eco de la distinción entre términos T-teóricos y términos T-no-teóricos, propia de los filósofos de la ciencia estructuralistas (véase KUHN, 1993: p. 333). Considera que se puede establecer una jerarquía de términos teóricos dentro del lenguaje de la ciencia, correspondiéndose cada nivel con una teoría histórica particular. 5) Al término de este proceso, podremos constatar la existencia de fallos de comunicación entre el grupo que ha renovado parte de su léxico, y el resto de los integrantes de la comunidad científica, que no se han visto en la necesidad de asumir tales cambios. Con el tiempo, van surgiendo nuevos campos de estudio y áreas profesionales, derivando en el alto grado de especialización y diversificación del saber a que hemos llegado en nuestra época. Nadie puede pretender, en la actualidad, abarcar todos los campos del saber científico. Es más, en beneficio del progreso, cada uno deberá elegir aquella especialidad que esté más en consonancia con su vocación y sus capacidades. El ensanchamiento progresivo de los límites del conocimiento humano va acompañado de una creciente división del trabajo científico.

Para Kuhn, las revoluciones son producto del intento de lograr que el léxico de la ciencia vuelva a satisfacer el “principio de no solapamiento”, una vez descubierto que ha sido violado. Según este principio, no puede darse el caso de que dentro de un mismo léxico dos términos de clase se solapen en sus referentes, a menos que se dé entre ellos una relación de especie a género (Kuhn, 1994: p. 30). La violación de esta condición obliga a una reelaboración de parte de la taxonomía léxica, mediante la incorporación de nuevos grupos de términos y el establecimiento de relaciones de clase diferentes. Utilizaré dos ejemplos del propio Kuhn. El primero es el descubrimiento de que no le es esencial al agua su estado líquido, pudiendo presentarse en estado sólido o como vapor (Kuhn, 1990: pp. 311-313). Este hallazgo condujo al abandono de las viejas taxonomías y a la aparición de un nuevo sistema taxonómico en el que se contemplan los tres estados de agregación del agua. El principio de no solapamiento prohibía la coexistencia de dos léxicos diferentes solapados en un mismo referente, la sustancia agua. Cuando un referente se halla en la región de solapamiento entre dos teorías alternativas, está sujeto a leyes incompatibles. La ciencia, en virtud de su ideal de coherencia, debe eliminar esta incompatibilidad, lo cual puede lograrse por alguna de estas dos vías: o bien se abandona una de las teorías, en favor de aquella que ha demostrado tener mayor éxito en la explicación de los fenómenos considerados relevantes; o bien la comunidad se divide para dar lugar a dos especialidades con léxicos diferentes. Esto mismo valdría para el ejemplo de ciencia ficción de Putnam (Ibíd., pp. 309-311), en el que se contemplan dos sustancias distintas como referentes del término ‘agua’, según éste sea utilizado por los habitantes de la Tierra o por los pobladores de la Tierra Gemela. El descubrimiento de muestras de una sustancia similar en todos sus aspectos al agua, pero con diferente composición química, obligaría a revisar la tabla de los elementos químicos. Las propiedades superficiales del agua tienen el mismo rango ontológico que las esenciales; ambas son necesarias. La teoría química que identifica el agua como H_2O , tiene que poder explicar sus cualidades superficiales; de lo contrario, no sería aceptable.

En “El camino desde «La Estructura»”, Kuhn parece distinguir tres niveles distintos dentro de un léxico, atendiendo a los distintos grados de convergencia o divergencia léxica entre individuos pertenecientes a la misma o diferente comunidad (véanse las pp. 41-45 de este artículo). En el nivel más básico se hallan aquellos aspectos, en su mayoría inconscientes, que están biológicamente determinados y que constituyen la expresión de una filogenia común: “Subyaciendo a todos estos procesos de diferenciación y cambio, debe haber, por supuesto, algo permanente, fijo y estable. Pero, como las “Ding an sich” de Kant, este algo es inefable, indescriptible, indiscutible” (KUHN, 1994: p. 45). En realidad, la versión local de la inconmensurabilidad ha dejado abierta esta posibilidad, ya que se admite que toda revisión conceptual tiene un alcance limitado y no constituye nunca una amenaza a la

totalidad de nuestro sistema de creencias. Da la impresión de que se está volviendo a la idea neopositivista de la existencia de cierto conocimiento observacional básico y estable, inmune al cambio revolucionario. A continuación, en un segundo nivel, se sitúan todos los conocimientos adquiridos mediante socialización, y que nos convierten en miembros de una cultura o grupo determinado. Es aquí donde cabe empezar a hablar de divergencia léxica, donde se libran los debates en torno a la inconmensurabilidad y la comunicación entre lenguajes y teorías diferentes. Con el paso de un lenguaje a otro, no se preservan las relaciones estructurales entre las palabras. Por último, el tercer nivel de variabilidad léxica se deriva de las diferencias interindividuales: dentro de una misma comunidad, cada individuo utiliza un léxico distinto y posee diferentes expectativas acerca de los fenómenos. Lo que permite que puedan entenderse y comunicarse entre sí es la idéntica estructura conceptual y taxonómica de tales léxicos, esto es, que se refieran a los mismos objetos y situaciones. La socialización consiste en la interiorización de ciertas estructuras provenientes de nuestro entorno físico y social particular. Estas estructuras son las condiciones de posibilidad de nuestra experiencia; pero, a diferencia de las categorías kantianas, son relativas a cada cultura y cambian a medida que cambian nuestros conocimientos (Ibid.: p. 45 y KUHN, 1993: pp. 331-332). El léxico de cada individuo es inseparable de su nicho: si, por un lado, su estructura conceptual lo vincula a un grupo o comunidad determinados, por el otro, los contenidos de esas estructuras reflejan una perspectiva particular acerca del mundo.

Kuhn no parece concederle demasiada importancia al problema de la unificación de las ciencias. Con la limitación del alcance de la inconmensurabilidad, todavía puede seguir manteniéndose la existencia de cierta continuidad entre sus etapas de desarrollo. Por otra parte, el que haya recurrido de forma tan insistente a metáforas biológicas para describir el proceso de constitución de las especialidades que integran la ciencia, nos sugiere que la imagen del árbol, utilizada ya en el siglo XVII por Descartes, sigue estando presente en su mente: la filosofía es el tronco del que han brotado los distintos saberes, de modo directo, o evolucionando a partir de alguna de las ciencias derivadas de ella. Con la mejora de nuestras herramientas conceptuales, el horizonte de la ciencia se ha ensanchado considerablemente, haciéndose necesario repartir nuestros esfuerzos en un radio cada vez amplio. Como consecuencia, se ha visto mermada la comunicación entre comunidades de científicos cuyas investigaciones se proyectan sobre campos diferentes.

3. Comparación con la noción de especialización estructuralista

La pretensión de los filósofos de la ciencia estructuralistas es dar cuenta de la constitución del conocimiento científico, integrando en sus planteamientos las perspectivas estática y dinámica. En otras palabras, se

proponen llevar a cabo un análisis y reconstrucción de las teorías científicas, mediante el uso de un aparato formal adecuado. Recuperan parte del legado neopositivista e historicista, haciéndolos confluir en sus propios planteamientos.

En cuanto a la tesis de la inconmensurabilidad, se pronuncian a favor de una versión local, en consonancia con la defensa de un holismo local. Para Ulises Moulines, el significado de un término depende de tres tipos de totalidades diferentes: complejos sistemáticos (formados por los modelos en que aparece determinado dicho término), complejos conceptuales (unidades conceptuales superiores de las que forma parte el término en cuestión) y complejos teóricos (conexiones interteóricas en las que está implicado el término). Este holismo es moderado, comparado con el holismo global, que hace depender el significado de un término de la totalidad de sistemas, términos y teorías pertenecientes a una disciplina concreta (MOULINES, 1991: pp. 197-199). Los filósofos de la ciencia estructuralistas admiten la existencia de teorías inconmensurables, si bien para ellos dicha inconmensurabilidad es local, afectando sólo a cierta región de ambos sistemas teóricos. Hasta aquí las confluencias entre Kuhn y los estructuralistas son significativas.

En cambio, en la caracterización de la noción de especialización, los enfoques son diferentes. Kuhn retoma la interpretación tradicional del concepto e intenta probar que la inconmensurabilidad es una consecuencia del proceso de especialización a que se ve sometido el conocimiento desde sus inicios. En ocasiones, parece identificar la relación de especialización con la de teorización: cuando surge una especialidad dentro de la ciencia, accedemos a un nivel superior de teorización. La nueva teoría dispondrá de un vocabulario antecedente (términos no-teóricos), cuyo significado depende de teorías preexistentes, situadas en algún nivel inferior; y una serie de términos característicos (vocabulario teórico). En cambio, para los estructuralistas, la relación de especialización se da entre las distintas especialidades surgidas a partir de un núcleo teórico, cuando se limita el alcance de la(s) ley(es) fundamental(es), mediante la formulación de alguna ley especial (MOULINES, 1991: pp. 244-246, 248-251). Por ejemplo, en la axiomatización de la Mecánica de Choque, sólo tenemos una ley fundamental -la ley de la conservación del momento. Si, en lugar de tomar en consideración todo el dominio de aplicación de esta ley, nos quedamos sólo con los choques elásticos, tendremos que además de la ley mencionada se cumple una ley especial: la ley de la conservación de la energía cinética. Las distintas especialidades obtenidas mediante la restricción del alcance de las leyes fundamentales de una teoría a ciertos contextos determinados, comparten un núcleo común (MOULINES, 1991: pp. 288-292). Este núcleo, también llamado "elemento modeloteórico", está constituido por ciertos conceptos básicos y las leyes fundamentales de la teoría inicial. El proceso de

especialización sucesiva de un elemento modeloteórico básico nos permite obtener enunciados de mayor contenido empírico y fuerza predictiva.

Así caracterizado, el concepto de especialización estructuralista no guarda relación alguna con la noción de incommensurabilidad. En el marco de la Concepción Estructuralista podríamos decir que se da incommensurabilidad entre dos teorías científicas diferentes, pero sería un error calificar como incommensurables a dos especialidades obtenidas en el seno de una misma teoría. Para que pueda hablarse de incommensurabilidad, las teorías comparadas tienen que tener núcleos diferentes.

Para terminar, diré que los conceptos de especialización de Kuhn y estructuralista no sólo me parecen diferentes, sino también compatibles: el primero tiene que ver con la progresiva fragmentación de la ciencia, a medida que aumenta su fuerza cognitiva; el segundo permite representar una teoría científica como una red cuyos nudos tienen en común un elemento modeloteórico. El marco respecto al cual se define cada uno de estos dos conceptos es diferente: Kuhn habla de la especialización del saber científico en general; en cambio, los filósofos de la ciencia estructuralistas hablan de las distintas especialidades derivadas de una misma teoría.

BIBLIOGRAFÍA

- Kuhn, T.S.: "Commensurability, Comparability, Communicability", en Asquith, P. D.- Nickles, T. (eds.): *Proceedings of the 1982 biennial meeting of the PSA*, 2. Michigan, Philosophy of Science Association, East Lansing, 1983, pp. 669-688 y 712-716. (Vers. cast.: "Commensurabilidad, comparabilidad y comunicabilidad", en Kuhn, T. S.: *¿Qué son las revoluciones científicas? y otros ensayos*. Paidós, Barcelona, 1989, pp. 95-135.)
- "Rationality and Theory Choice", *Journal of Philosophy*, 80 (1983): 563-570. (Vers. Cast.: "Racionalidad y elección de teorías", en Kuhn, T. S.: *¿Qué son las revoluciones científicas...?*, op. cit. pp. 137-151.)
- "What are Scientific Revolutions?", en Kruger, L.- Daston, L. J.- Heidelber, M. (eds.): *The Probabilistic Revolution, 1: Ideas in History*. Cambridge (Mass.), M.I.T. Press, 1987. (Vers. cast.: "¿Qué son las revoluciones científicas?", en Kuhn, T. S.: *¿Qué son las revoluciones científicas? y otros ensayos*. Paidós, Barcelona, 1989, pp. 55-93.)
- "Dubbing and Redubbing: The Vulnerability of Rigid Designation", en Wade Savage, C. (ed.): *Scientific Theories*. Minnesota Studies in the Philosophy of Science 14. University of Minnesota Press, Minneapolis, 1990, pp. 298-318.
- "The Road Since Structure", en Fine, A.- Forbes, M.- Wessels, L.: *Proceedings of the 1990 biennial meeting of the PSA* 2. Philosophy of Science Association. East Lansing, Michigan, 1991, pp. 3-13. (Vers. cast.: "El camino desde «La Estructura»", *Arbor*, 148, n. 583 (1994), 27-46.)
- "Afterwords", en: Horwich, P.: *World Changes. Thomas Kuhn and the Nature of Science*. The MIT Press, Cambridge, 1993, pp. 311-341.
- Moulines, C.U.: *Pluralidad y recursión*. Estudios epistemológicos. Alianza, Madrid, 1991.

Patricia Rodríguez Fernández
Universidad de Santiago de Compostela
e-mail: patri-rf@teleline.es



CONSTRUCTIVISMO SIN EMPIRISMO: Crítica a la teoría sensualista de van Fraassen*

EKAI TXAPARTEGI
Universidad del País Vasco

ABSTRACT: The main thesis of the communication is that van Fraassen's account implicitly succumbs to a naive realism. The constructive empiricism characterised by van Fraassen introduces a conceptual distinction between 'seeing' and 'seeing that' where there is, like I will argue, a unidirectional relation of underdetermination by the former to the latter. My hypothesis is that a better account is provided by Cassirer's representational idealism. Cassirer holds an analogous distinction between 'presented' and 'represented', although with a basic difference: Cassirer does not accept 'present' without 'represent'. The complexity introduced by Cassirer shows the bidirectionality to the above mentioned epistemic process of construction of 'observables'.

KEYWORDS: Seeing, observability, representational view, ontology, van Fraassen, Cassirer.

Introducción

¿Cuándo estamos objetivamente justificados a aceptar una teoría? Cuando hace veinte años van Fraassen lanzó esta pregunta la solucionó diciendo que una teoría puede ser aceptada si es empíricamente adecuada, esto es, si las subestructuras observables de la teoría son verdaderas. El requisito para la aceptación de la teoría, según van Fraassen, no es pues la verdad de toda la teoría sino tan sólo la verdad de una de sus partes, la parte observable. La otra parte, las estructuras matemáticas que satisfacen la teoría, esto es, los modelos, no tienen un valor de verdad si no es por derivación, si no es a través de la hipótesis teórica que enlaza la parte formal de la teoría con los sistemas empíricos observables.

¿Debemos creer la teoría que aceptamos? La respuesta de van Fraassen a esta segunda pregunta viene determinada por la partición de la

* Esta comunicación ha podido ser realizada en parte gracias a la financiación del Gobierno Vasco por medio de la Beca para la Formación de Investigadores concedida por el Departamento de Educación, Universidades e Investigación del Gobierno Vasco.

teoría que basándonos en él acabamos de trazar. Como hemos dicho para aceptar una teoría es suficiente con aceptar la verdad de una de sus partes. En cambio, para creer la letra de la teoría, y no simplemente aceptarla o utilizarla, no es suficiente con la adecuación empírica de una de sus partes. El compromiso ontológico que se adquiere al aceptar una teoría no va más allá del compromiso con los observables. Y puede que ni siquiera llegue hasta ahí, ya que como explícitamente dice "observable y existe" no se implican mutuamente. Por lo tanto, deberíamos distinguir, siguiendo el criterio de van Fraassen, entre las teorías y las distintas actitudes proposicionales (de creencia, de duda etc.) que mantenemos sobre ellas.

Mi tesis principal es que el antirealismo de van Fraassen viene impregnado, por contradictorio que pudiera parecer, por una teoría sensualista marcadamente realista que no justifica suficientemente. Mantendré que van Fraassen comparte la intuición básica del realismo clásico que sostiene que el origen causal de lo que conocemos sobre el mundo es el mundo y no el sujeto. En este sentido, su conocida proclama de salvar los fenómenos significa, tal y como trataré de argumentar, salvar el 'núcleo originario' sobre la cual se construye el edificio de la ciencia. Lo que pretendo hacer a lo largo de la comunicación es mostrar la ilegitimidad de ese 'núcleo originario' que introduce van Fraassen, basándome sobre todo en lo que interpreto es su teoría sensualista. Mostraré asimismo que el idealismo representacional de Cassirer puede ofrecer a la cuestión acerca del estatus ontológico de los datos de los sentidos una salida que sea más plausible y menos comprometida metafísicamente, que los postulados empiristas en los que se basa van Fraassen. Al fin y al cabo de lo que se trata es de ahondar un poco más en el viejo dilema entre el empirismo y el idealismo, pero de una manera muy específica. Me preguntaré: ¿está justificado el empirismo de van Fraassen?

La teoría sensualista de van Fraassen: 'Ver' vs 'Ver que'

Una cuestión filosóficamente interesante de la teoría sensualista es saber cómo se configuran los datos de los sentidos. Cifrándonos al empirismo constructivista postulada por Van Fraassen, acabamos de decir que la hibridez ontológica de las teorías científicas le permite salvar la objetividad y la inteligibilidad de nuestras actitudes doxásticas salvando los observables. Ya que, como hemos señalado más arriba, el compromiso ontológico que se adquiere al aceptar una teoría no va más allá del compromiso con los observables. En otras palabras, según van Fraassen, lo que estaríamos justificados a creer acerca de lo que en el mundo haya cuando se acepta una teoría no es más que aquello que directamente se puede observar. Por lo tanto la cuestión interesante de la teoría sensualista de van Fraassen es dilucidar

cómo cree él que se configuran los 'observables', exactamente, *¿qué grado de independencia teórica otorga van Fraassen a lo que llama 'observables'?*¹.

En este sentido, en contra de lo que voy a decir en esta comunicación, parece que Van Fraassen tampoco confiere demasiada independencia a lo que considera observable. En una primera lectura pudiera parecer que la acusación de que van Fraassen sostiene una teoría sensualista de corte realista carece de fundamento. Rastreando sus escritos con detenimiento se encuentra que *la dependencia teórica de lo observable distingue por lo menos dos niveles*, en otras palabras, que la carga pragmática que contiene la teoría impregna a lo observable por lo menos por dos vías conceptualmente distintas:

- a) *La pluralidad de los observables*: La misma visión puede conducir a la construcción de diversos observables, por el hecho de que, según van Fraassen, lo que vemos, esto es, aquello que conceptualmente vemos, el susodicho 'ver que', viene ciertamente determinado por el marco cognoscitivo. De esta manera, una pelota de tenis es un observable pero siempre dentro de un cierto contexto conceptual. 'Ver' una pelota de tenis no es un mero hecho perceptivo sino que conlleva una dirección conformadora que en otro contexto podría dar como resultado otra descripción de lo visto, otros observables, quizá inconmensurables entre sí. Con ello se mantiene que efectivamente pueden construirse una pluralidad de observables desde los datos crudos de los sentidos según los criterios pragmáticos que se compartan en cada contexto cultural.
- b) *El carácter funcional o relativo de los observables*: En la descripción que hace de la construcción de las teorías [Van Fraassen, 1989, 228-230] admite que son las teorías las que señalan lo que cuenta como observable para dicha teoría. Una cuestión interesante desde esta perspectiva, que van Fraassen no trata, sería dirimir si lo que es observable para una teoría puede ser no-observable (y no simplemente trivialmente observable) para otra teoría. Pero en todo caso, al admitir que es la teoría la que en cada caso determina qué es lo que cuenta como observable, van Fraassen abre la puerta para una primera aproximación funcional de lo observable, ya que lo que cuenta como observable, no es absoluto sino relativo a cierta teoría.

La conclusión más común que se extrae del análisis hecho a estas tesis epistémicas de van Fraassen, es que la dinámica entre la teoría y los fenómenos observables puede considerarse en este primer sentido bidireccional: por un lado la teoría se levanta sobre la libertad creadora que ofrecen los fenómenos observables que hay que salvar; por otro lado, tal y

¹ Sobre el carácter pasivo o dinámico de la relación entre la teorías y lo que en la teoría cuenta como observable véase, Ibarra-Morman 73-90 y A. R. Pérez Ransanz '*La concepción semántica de las teorías y el debate sobre el realismo científico*'.

como lo describen las condiciones a) y b), los fenómenos observables vienen de cierta medida determinados por la teoría.

La intención de esta comunicación es mostrar que esta bidireccionalidad aparente, que supone una clara afinidad con la tesis de la teórico-dependencia de los observables, mantiene un recodo realista que remite a una base 'no-conformada' teóricamente con lo que la bidireccionalidad se localizaría en un segundo nivel y no afecta a todo el sistema. Esa base virginal, no contaminada por la teoría, se sustenta, como trataré de argumentar, en su teoría sensualista. En la fase constructiva trataré de dar una alternativa plausible a la concepción sensualista del empirismo de van Fraassen que consiste en una concepción representacional de todo hecho cognoscitivo, incluso al nivel de las percepciones sensoriales, basada sobre todo en la teoría anti-sensualista del idealismo cassireriano.

Descartada definitivamente la teoría correspondencialista de la verdad con un mundo ya partido de antemano el anhelo realista puede transformarse en ciertos casos en una especie de empirismo donde la objetividad que garantizaba el 'Ser' se transforma en una objetividad garantizada por el 'Ver'. Según este modo de entender el conocimiento, que creo comparte *el empirismo constructivista de van Fraassen*, 'Ver' sustituye a 'Ser' en la función de amarre del pensamiento. La tarea que se impone a partir de aquí es definir el 'ver' según la teoría sensualista que se maneje. Van Fraassen, creo yo, nunca ha elaborado explícitamente una teoría sensualista que defina e interprete los ítems percibidos. A pesar de ello su empirismo basado en los 'observables' da muchas pistas de lo que pueda ser para van Fraassen lo meramente percibido. Debido a este hecho el punto de partida de mi análisis serán los observables.

Van Fraassen mantiene una *idea muy antropocéntrica de lo observable*. Observable, contrariamente según él al concepto de existencia, es un concepto completamente antropocéntrico: "*El organismo humano (...) tiene ciertas limitaciones inherentes, que serán descritas en detalle por la física y la biología últimas. Son estas limitaciones a las que el 'able' en observable se refiere: nuestras limitaciones qua seres humanos*"² [van Fraassen, 1980, 34]. Ésta es según la interpretación de van Fraassen la forma más básica de las limitaciones epistemológicas del género humano. De acuerdo con este 'able' el dominio de las 'cosas del mundo' pueden partirse en dos: las cosas que pueden

² Churchland introduce una triple distinción entre "1.-cosas que han sido observadas por algunos humanos, 2.- cosas que pueden ser observadas por los humanos y 3.- cosas que no pueden ser observadas por los humanos" con el propósito de reforzar la tesis de la completud ontológica de la teoría. Según Churchland no hay una distinción ontológica entre los observables y los no observables "*In fact, their epistemic situation (el de los observables) is not superior. They are exactly as dubious as their nonobservational cousins. Their causal history is different (they are occasioned by activity in the sensory pathways) but the ontology they presuppose enjoys no privilege or special credibility*"[41].

ser observadas y las cosas que no pueden ser observadas por la especie humana. Lo observado por consiguiente es una subclase de los observables que son, a su vez, relativos a la capacidad epistémica humana. Si vemos una pelota de tenis, siguiendo con el ejemplo, es porque la pelota de tenis es, de hecho, un observable según las capacidades epistémicas de la especie a la que pertenecemos. Lo que cuenta en última instancia son pues las capacidades y las limitaciones epistémicas de cada especie biológica, en sus propias palabras, su rango de evidencia accesible. Él mismo concluye que "*lo que cuenta como un fenómeno observable está en función de lo que es la comunidad epistemológica (que observable es observable-para-nosotros)*" [van Fraassen, 1980, 36].

Pero bien al contrario mantiene también *una idea muy poco intencional de la observación*. Para ello traza una distinción conceptual entre 'observar' (o1) y 'observar que' (o2) [van Fraassen, 1980, 32]. Los observables se corresponden a 'lo que observamos' (o2), esto es, al resultado de la actividad observacional pasada por el tamiz de las condiciones a y b anteriormente señaladas y que como hemos advertido antes empañan de teoría la actividad puramente observacional (o1). Los observables (o2) son, en definitiva, la suma de la observación (o1) y de los aspectos pragmáticos. Los observables remiten de este modo a una fase anterior donde la observación (o1) se mantiene invariable a través de los diversos marcos conceptuales que se elijan. Los observables son contruidos a partir de los ítems perceptuales. Y estos ítems perceptuales, fruto del o1, son, por un lado, el límite epistémico más básico de la especie humana y, por otro, algo que se mantiene independiente de su intencionalidad o de los criterios pragmáticos subyacentes a los marcos conceptuales que se comparten a distintos niveles.

La tesis de la independencia del 'ver' (o1) se sostiene sobre la tesis de la invariabilidad del 'ver' (o1). Esta última tesis sostiene que aunque utilicemos teorías científicas distintas vemos (o1), en el sentido básico de la palabra, lo mismo. Es una tesis que aunque la defiende con contundencia no la argumenta, simplemente la constata: "*decir que él (el hombre de la Edad de Piedra encontrado en las Filipinas) no ve (o1) lo mismo es simplemente tonto*" [van Fraassen, 1980, 32]. En cambio mi hipótesis es que puede mantenerse la imposibilidad de discernir si efectivamente se ve (o1) lo mismo o no independientemente de los esquemas conceptuales, y que por lo tanto si se quiere mantener que se ve (o1) lo mismo sin caer en un círculo vicioso debe ser explicado sin apelar a la supuesta independencia del 'ver' (o1) con respecto al 'ver que' (o1).

Para dar cuenta de la supuesta invariabilidad del 'ver' (o1) prefiero empezar por analizar los casos en los que van Fraassen otorga *variabilidad a 'ver' (o1)*³. Siguiendo a van Fraassen, o1 puede ser variable o puede

mantenerse invariable, depende de si la especie biológica se mantiene o no invariable. Van Fraassen lanza dos hipótesis para justificar la variabilidad de o1: una mutación genética o la incrementación de la comunidad epistémica por medio de una decisión ideológica o moral. Según esta hipótesis lo que varía es la capacidad de los órganos perceptores que sufrirían una mutación o que incrementarían su potencial al tomar como percepciones válidas también las percepciones de aquellos grupos epistémicos que tuvieran órganos perceptores diferentes⁴. En tal supuesto comunidades epistémicas diferentes no verían lo mismo, no manejarían los mismos datos y la inconmensurabilidad posible se podría incluso extender a la estructura de los datos brutos.

Si se postulan dos comunidades epistémicas diferentes que difieren hasta en la propia apercepción, surge la pregunta acerca de *cómo se comparan los distintos 'veres'*, para determinar si realmente son distintos, sin apelar a aquello que es visto ya que lo que se cuestiona no es el 'ver que' (o2) sino el mismo 'ver' (o1): En este caso los observables no pueden ser comparados de ninguna de las maneras, ni en el sentido realista, ni en el instrumentalista, ni en el conductista⁵ (tal y como pretende van Fraassen cuando en un momento dado afirma que los hombres de Filipinas también reaccionan ante un accidente de coches), ya que hablamos de lo que perceptualmente, en el sentido pre-sintético de la palabra, registran comunidades epistémicas diferentes como lo pueden ser la especie humana y la de los extraterrestres sin mencionar en cualquier caso ni qué ni cómo lo configuran. La cuestión reside, según lo interpreto yo, en la simple idea de que si se dice que dos comunidades epistémicas ven lo mismo, debe haber también algún modo de determinar si efectivamente se trata de los mismos 'veres', sin hacer referencia alguna a 'lo visto'.

Según mi interpretación, de las hipótesis propuestas por van Fraassen en los dos ejemplos se puede inferir que de hecho sí hay algo que se puede

⁴ Debo señalar que van Fraassen al lanzar esta hipótesis supone que habrá algún modo de compatibilizar aquellas percepciones del grupo epistémico que se pretende integrar en nuestra comunidad epistémica. Hipótesis que contradice su suposición principal que es la de la inconmensurabilidad radical entre grupos epistémicos diferentes. Pero en pro de la argumentación supondremos que puede haber tal grupo epistémico de extraterrestres que sea diferente aunque compatible con el nuestro.

⁵ Quine asimismo en su libro de 1974 *Las raíces de la Referencia* también mantuvo, al igual que van Fraassen, la validez de criterios conductistas en la comparación de los diversos ítems registrados en los órganos perceptuales. Pero también al igual que van Fraassen mantuvo su posición sin justificarla. *"Los receptores sensibles operan en el nivel de la recepción, mientras que la Gestalt opera en el nivel de la percepción. El viejo antagonismo se debió a la deformada parcialidad del epistemólogo respecto de la recepción, a la que concebía como necesitada de consciencia, cuando ésta corresponde sólo a la percepción. Le recepción es manifiestamente física. Pero también la percepción es accesible a criterios conductistas, pese a todas sus resonancias mentalistas."* Con este giro cierra Quine su tratamiento de lo registrado de modo pre-consciente y comparado de manera conductista.

comparar entre dos comunidades epistémicas que no 'ven' lo mismo: *los órganos perceptores*, por los cuales se diferencian⁶. Es bastante plausible que los humanos y los delfines no vemos lo mismo puesto que los órganos perceptuales no son compartidos. Pero esta definición también puede ser problemática ya que 'ver' se identifica con el registro primitivo que se da en los órganos perceptores: 'ver' es lo que se registra, bajo el supuesto de que lo que se registra es lo mismo entre los que comparten los mismos órganos perceptuales y distinto entre los que no los comparten. Suponer que lo que se registra es diferente según los diferentes órganos perceptuales es aceptar que los órganos perceptuales de alguna manera conforman aquello que perciben. Aceptando esto último podemos definir a una comunidad epistémica diferenciada como aquella comunidad que cuenta con órganos perceptivos propios. Así se constata que diferentes comunidades epistémicas no 'ven' lo mismo si no comparten los mismos órganos perceptuales. Por lo tanto siguiendo a van Fraassen hay efectivamente un criterio bastante plausible de demarcación de comunidades epistémicas que no 'ven' lo mismo, en el sentido fuerte de la palabra.

La tercera conclusión que se refiere a la caracterización de la teoría sensualista de van Fraassen es que aún habiendo en la observación un ámbito independiente de criterios pragmáticos, no es un ámbito completamente independiente ya que viene determinado por los órganos perceptivos con los que se realiza dicha percepción. Por todo ello se puede mantener que 'ver' es 'órgano-dependiente' y que 'ver que' es a su vez órgano-dependiente y contexto-dependiente, según la distinción marcada por van Fraassen.

El idealismo representacional de van Fraassen

Tal y como se formula la propuesta realizada por van Fraassen hablar de la varianza de 'ver' supone forzosamente hablar de la comparabilidad de apercepciones desnudas de todo marco conceptual. Así lleva el problema de si se ve lo mismo al debate de la comparación de los diversos sistemas perceptivos. A pesar de ello, de este modo, creo yo, no evita un debate que es previo, esto es, el debate de fondo acerca de qué es lo que se considera como 'órganos perceptuales', si son simplemente los órganos propiamente dichos o el sistema perceptual entero creado en torno a ellos que podría llevarnos a *negar la posibilidad misma de distinguir entre 'ver' y 'ver que'*. Distinción que

⁶ Para comparar aquello que se registra Quine también recurre al análisis de los órganos perceptuales, o siendo más exactos, al análisis de lo que la ciencia dice acerca de los órganos perceptuales: "*la ciencia misma enseña que no hay visión intuitiva, que la única información que desde los objetos externos puede alcanzar nuestras superficies sensoriales tiene que estar limitada a proyecciones ópticas bidimensionales, varios impactos de ondas de aire en los tímpanos, algunas reacciones gaseosas en las fosas nasales y unas pocas cosillas más análogas.*"

parece simplificaba en demasía el posible funcionamiento real de nuestro sistema perceptual.

En este sentido puede ser esclarecedor el debate que mantuvo Cassirer a propósito de las tesis sensualistas que Marc-Wogau atribuyó a su libro *'La Filosofía de las Formas Simbólicas'*. El idealismo representacional de Cassirer mantiene que la coerción constitutiva del símbolo -del representante- opera desde el nivel mismo de la percepción sensorial. Análogamente a la distinción de van Fraassen entre 'ver' y 'ver que' Cassirer distingue entre 'presentar' y 'representar', para borrar los límites de tal distinción: *"Lo que Marc-Wogau objeta, sin embargo, es que en mi propia teoría el sensualismo no está todavía superado y eliminado por completo. Porque también en ella se distingue entre un contenido dado aquí y ahora y aquello que dicho contenido representa y hace presente, entre 'presencia' y 'representación'. Pero, ¿déjase efectuar semejante distinción, ya que una separación real entre los dos elementos no sería posible, ya que un fenómeno no es nunca un dato meramente sensible que sólo se experimente como tal, sino que se junta con otros en un todo y 'apunta' indirectamente a dicho todo, 'representándolo' o 'presentándolo'?"* [Cassirer, 1956, 195].

Cuando el *sensualismo ingenuo* de van Fraassen afirma que la comunidad epistémica que comparte ciertos órganos perceptuales ve, o podría potencialmente llegar a ver, lo mismo, se está haciendo una afirmación sobre el carácter igualmente conformador de los órganos perceptuales, donde se manifiesta la realidad en un grado todavía no conformado por el entendimiento subjetivo. La pregunta que Cassirer lanzaría en este momento a van Fraassen sería: *"¿Existe, pues, en general, una materia de lo real, anterior a toda conformación e independiente de ella?"* Interpreto que el empirismo de van Fraassen pretende ser una respuesta afirmativa a esta pregunta basada en una muy débil concepción de los órganos sensoriales, ya que toma como punto de partida el determinado complejo de impresiones sensibles que, en cuanto datos elementales, están a su entender directamente dados, aunque efectivamente puedan después desarrollarse conceptualmente en distintas direcciones.

Conclusión

Recapitulando, van Fraassen ha mantenido que los observables estaban doblemente infradeterminados por el 'ver'. Los observables, 'aquello' que se podía ver, era a su vez órgano-dependiente y contexto-dependiente. 'Ver' en cambio era solamente órgano-dependiente. Pero esta distinción no es una distinción ni absoluta ni una distinción que sirva para aclarar mucho pues nada hay, parece ser, que pueda ser señalado como 'ver' en el sentido virginal de ser solamente órgano-dependiente. *"Coincido por mi parte con esta concepción de la psicología moderna en dos puntos esenciales. Insisto de la manera más categórica en que la 'mera' percepción, o sea la percepción en*

cierto modo desnuda, que estaría libre de toda función de representación, no es un fenómeno que nos esté dado directamente en nuestra 'actitud natural'(...)"[Cassirer, 1956, 199] . Como adelantó Churchland apenas se publicó el libro de van Fraassen "Inflationary metaphysics and 'pragmatic' decisions begin, it seems, as soon as we open our eyes" [Churchland, 1985, 42].

El punto de partida de toda teoría representacional interesante es la interdependencia esencial, no excepcional, entre 'ver' y 'ver que', hasta llegar a no poder distinguirse. En este sentido y retomando la argumentación que utilizó van Fraassen para introducir tal distinción, quizá sea tonto decir que cuando Einstein y Tolomeo miraban al cielo no veían lo mismo, pero ese 'ver lo mismo' debe ser explicado sin apelar a la distinción entre 'ver' y 'ver que'. Ya que parece ser que no debemos aceptar ninguna clase de actividad cognoscitiva, de actividad representacional, que mantenga su espacio de acción libre de toda configuración del sistema. La bidireccionalidad postulada entre los fenómenos observables y la teoría alcanza de esta manera también el nivel básico de los datos elementales que van Fraassen pretendía resguardar en esa caja negra que es el empirismo sensualista.

Una vez aceptada la bidireccionalidad de la función de representación que afecta a todo el sistema el siguiente paso es construir una distinción entre los dos polos de la teoría o de cualquier sistema representacional verdaderamente funcional o relativa a éstas. El estudio de dicha dinámica representacional y la aceptación de la tesis que postula la constitución interna de la ontología pueda quizá arrojar nueva luz hacia aspectos ontológicos que en la concepción de van Fraassen permanecían confinados en la esfera de las creencias metafísicas.

BIBLIOGRAFÍA

- Cassirer, E. (1956): *Esencia y Efecto del Concepto de Símbolo*, México: FCE 1989.
Churchland, P.M. Hooker, C.A. (ed): *Images of Science*, Chicago: Univ. of Chicago Press 1985.
Ibarra, A., Mormann, T.: *Representaciones en la Ciencia*, Barcelona: Ed. del Bronce 1997.
Moulines, U.: '¿Qué Clases de Cosas Hay?', en: Gómez Pin, V. (ed.), *Congreso Internacional de Ontología 1993*, Universidad Autónoma de Barcelona 1994.
Quine: *Las Raíces de la Referencia*, Madrid: Revista de Occidente, 1977
van Fraassen, B.C., (1980): *La Imagen Científica*, Buenos Aires, México: Paidós, 1996.
van Fraassen, B.C.: *Laws and Symmetry*, Oxford: Clarendon Press°, 1989.

*Ekai Txapartegi,
Universidad del País Vasco (UPV/EHU)
e-mail: ylbtxzue@sf.ehu.es*



EL “PRINCIPIO DE FAMILIARIDAD” ¿EXISTE UN ARGUMENTO RUSSELLIANO A SU FAVOR?

JESÚS VEGA ENCABO
Universidad de Salamanca¹

ABSTRACT: In this paper, it is presented a Russellian argument for the defense of the Principle of Acquaintance” (PA) against some well-known criticisms. PA establishes some epistemic constraints derived from the cognitive relation of acquaintance at the level of thought and the understanding of language. By introducing a classical distinction between two different senses of “object of thought”, I review Russell’s reasons for accepting the principle. I argue that the existence of an object-directed, simple, cognitive relation such as acquaintance is decisive in the argument. Every reconstruction of the principle has to deal with the epistemological and metaphysical assumptions lying behind the relation of acquaintance.

KEYWORDS: Principio de familiaridad, objetos de pensamiento, pensamiento singular, saber qué, principio de discriminación, referencia.

1. La formulación más conocida del “principio de familiaridad” (PF) de Russell reza como sigue:

“Toda proposición que podemos entender debe componerse totalmente de constituyentes con los cuales estamos familiarizados” (Russell, 1910-11, 219; 1912, 58)

El principio, introducido ya en 1905 en su artículo “On denoting” como un resultado del análisis de las descripciones definidas, se aplica esencialmente en el nivel del pensamiento y, dada la concepción russelliana de las relaciones lenguaje-pensamiento, establece al mismo tiempo constricciones epistémicas sobre en qué consiste la comprensión de un lenguaje.

¹ Este trabajo fue redactado durante una estancia postdoctoral en Brown University financiada con una beca del Ministerio de Educación. Agradezco los comentarios de E. Sosa y F. Broncano.

2. Se ha generalizado la idea de que Russell no ofreció ningún argumento en defensa de *PF* e incluso que las razones aparentes que presenta le enredan en una petición de principio o en círculos viciosos (Hintikka, 1972; Sainsbury, 1986; Ackerman, 1987; Donnellan, 1990). No obstante, por un lado, Russell no pretende argumentar a favor de un principio que cree *intuitivamente* aceptable: “la verdad de este principio es evidente tan pronto como se comprende el principio” (Russell, 1910-11, 118). A pesar de ello, por otro lado, no deja de ser de interés preguntarse sobre si es posible reconstruir, a partir de las razones intuitivas que Russell esparce en sus escritos, un argumento en defensa de *PF* que no implique ninguna petición de principio ni circularidad.

3. Un breve comentario sobre las dos nociones fundamentales que entran en la formulación del principio, *constituyentes de una proposición y familiaridad*, ayuda a precisar su sentido.

La primera es una expresión extraña en boca de Russell después del año 1910, cuando rechaza explícitamente la existencia de proposiciones y abandona la teoría de que éstas sean los objetos de pensamiento y de juicio. Las proposiciones no son los objetos primarios del pensar, no son *aquello* que es creído, juzgado, deseado, etc. Russell creía tener dos razones para su exclusión: la primera apela a lo poco plausible que resulta hablar de “falsedades objetivas” o “hechos no reales”; la segunda se basa en un tratamiento de las cláusulas proposicionales como “símbolos incompletos” que no tienen significado por sí mismas. Su teoría del juicio y del pensar como un tipo de relación a un conjunto independiente de entidades se deriva fácilmente. El rechazo de la metafísica idealista permite asumir que “en todos los actos cognitivos, tales como creer, dudar, desconfiar, aprehender, percibir, imaginar, la mente tiene objetos distintos de sí misma con los que está en alguna de estas diferentes relaciones” (Russell, 1910, 173-4). Si juzgar o pensar es una relación de la mente a objetos, y estos objetos no son proposiciones, entonces parece ser una relación “relacionante” y múltiple a un conjunto de entidades. ¿Qué clase de entidades es aceptable como objetos no-mentales y suficientes para explicar la significatividad? La respuesta russelliana es que han de ser objetos de familiaridad, entre los que incluye los datos sensoriales, los universales y el yo². Así, para comprender una oración de un lenguaje (o tener un pensamiento), el sujeto debe “saber lo se quiere decir con las palabras”, lo que es equivalente para Russell a tener familiaridad con los objetos por los cuales están las palabras (Russell, 1913, 111). La comprensión (tener pensamientos) es una relación múltiple a un conjunto de objetos de familiaridad que representan los términos en los cuales pensamos *sobre* un objeto. Por un lado, estos términos describen aquello *sobre* lo que

² Russell cambiará de opinión sobre la inclusión del “yo” como objeto de familiaridad en 1913.

estamos pensando; por otro, *qué* —en qué términos— pensamos sobre ello³. Esta relación múltiple no es el resultado de relaciones duales de familiaridad con cada elemento: no se requiere que las familiaridades relevantes con cada término por separado se activen, pues no sería entonces más que un caso de familiaridad compleja, y Russell insiste en que la comprensión o el pensamiento no pueden ser entendidos como una relación de familiaridad. Es fácil comprobar, entonces, por qué sería preferible una formulación del *PF* que no mencionara los constituyentes de las proposiciones. El mismo Russell la introduce en sus escritos:

"En cualquier momento en que tiene lugar una relación de suponer o juzgar, los términos con los que está relacionada la mente que supone o juzga mediante la relación de suponer o juzgar deben ser términos con los cuales la mente en cuestión está familiarizada"
(Russell, 1910-11, 118).

¿En qué consiste esta relación de *familiaridad*? Dejando a un lado sus tintes metafóricos, he aquí una caracterización russelliana de la noción: *familiaridad* consiste en una relación cognitiva dual entre un sujeto y un objeto que proporciona un conocimiento directo (no-inferencial) e inmediato, no sujeto a duda (infalible e incorregible), un conocimiento completo y perfecto del objeto. Se trata de un tipo de relación entre sujeto y objeto inversa de la relación de *presentación*, de una forma de *aprehensión* por parte de la mente del objeto, de una relación en la que el sujeto llega a ser directamente *consciente* del objeto mismo. Aún los términos pueden dar lugar a ambigüedades, pues la *familiaridad* tendería a ser vista como un *acto* y como una *relación*. No sería arbitrario insistir en que la *familiaridad* es, estrictamente hablando, una *relación* que implica la externalidad de los objetos. si se quieren evitar ciertas confusiones relativas a las nociones de acto y objeto que generan posiciones idealistas. Con independencia de las dificultades para caracterizar de manera precisa la russelliana noción de familiaridad, queda en pie el hecho de que es un requisito para el pensar (para sostener relaciones cognitivas complejas como comprender, suponer, juzgar, etc.) el que la mente esté en una relación epistémica especial con los términos bajo los cuales pensamos sobre las cosas.

4. *PF* se traduce entonces en lo siguiente:

Toda proposición que podemos entender debe estar compuesta completamente de constituyentes con los cuales tenemos una relación epistémica A.

³ En *Principles of Mathematics*, Russell denominaba "términos" a las entidades referidas por las expresiones de un lenguaje. Cada término es *uno* y tiene "ser" (*being*). Los términos eran considerados igualmente como *objetos de pensamiento* y, por tanto, constituyentes de las proposiciones. Distinguía entonces dos tipos de términos: las cosas, designadas por los nombres propios; y los conceptos, denotados por otras palabras de un lenguaje (Russell, 1903, &47-48).

Sea p una proposición cualquiera; $C(p)$, el conjunto de sus constituyentes; ' sUp ', el hecho de que s entienda la proposición p , y A , este tipo de relación epistémica especial. PF afirma que para cada término x de una proposición cualquiera p que el sujeto s entiende, x pertenece a $C(p)$ sólo si el sujeto s está relacionado a x por la relación epistémica A .

$(PF) (s) (x) (p) L[(sUp \ \& \ x \in C(p)) \rightarrow sAx]$

La formulación alternativa, estrictamente russelliana, del principio rezaría del siguiente modo:

Cualquier relación cognitiva compleja de la que es capaz un sujeto implica sólo términos con los cuales el sujeto tiene una relación epistémica A.

Paralelamente, si R representa cualquier relación cognitiva compleja, PF se traduciría como:

$(PF') (s) (x) (r) L[sR(x_1... x_n) \rightarrow (sAx_1... sAx_n)]$

5. El principio pretende establecer condiciones epistémicas constitutivas de ser un objeto de pensamiento en general. Esta tesis requiere dos comentarios: la sustitución de la expresión "constituyente de la proposición" por "objeto de pensamiento" no resuelve ciertas ambigüedades; pero además, no convierte, en ningún caso, el principio en una perogrullada que afirme una equivalencia estricta entre objetos del pensamiento y de familiaridad.

La teoría russelliana de las descripciones definidas es decisiva a la hora de enfrentar el problema. ¿Cómo es posible que un sujeto s sea capaz de sostener pensamientos *sobre* el objeto significado bajo la descripción definida "el tal-y-cual" si no tiene ninguna relación de familiaridad con el objeto? La teoría busca resolver esta dificultad mediante un movimiento decisivo: la forma gramatical —argumenta Russell— disfraza y oculta en ocasiones la forma lógica. En el caso de las descripciones definidas, nos enfrentamos aparentemente a un dispositivo referencial que no denota necesariamente un objeto. No son expresiones referenciales genuinas, como los nombres propios lógicos, que garantizan la *existencia* del objeto. Para Russell, no son dispositivos aparentemente referenciales sólo las expresiones del tipo "el tal-y-cual", sino que además los nombres ordinarios no son más que descripciones definidas truncadas. La propuesta russelliana para su análisis es traducirlas al lenguaje de cuantificadores. Las descripciones definidas son expresiones cuantificadas que establecer que el predicado se satisface de manera única y que cualquiera que sea el objeto que satisface el predicado tiene cierta propiedad. "El tal-y-cual es F " es parafraseado como "hay exactamente un tal-y-cual y es F "⁴.

Aún así, un sujeto puede tener un pensamiento *sobre* un objeto cualquiera x sin que x mismo sea un "objeto de pensamiento". ¿Podemos tener pensamientos *sobre* objetos inexistentes? Es imposible tener una relación de

⁴ De manera más precisa: $(Ex) [(Px) \ \& \ (y)[Py \rightarrow (y=x)] \ \& \ Fx]$

familiaridad con un objeto que no existe. Russell insiste que todo objeto de familiaridad necesariamente existe (Donnellan, 1990). Por tanto, si el principio es válido, entonces o bien es imposible tener ningún tipo de relación cognitiva con pensamientos sobre inexistentes o bien el objeto como tal no es constituyente u objeto de ese pensamiento. Es evidente que comprendemos enunciados *sobre* objetos inexistentes. En consecuencia, es posible pensar sobre un objeto sin que este sea uno de los constituyentes del pensamiento.

Podría traerse a colación la distinción que Prior estableció en los usos de "objeto de pensamiento": el primero se centra en *lo que* pensamos; el segundo, en el *objeto sobre el cual* tenemos el pensamiento (Prior, 1971, 1-2). Es de justicia insistir en que el primer sentido no es equivalente, en la filosofía de Russell, al de proposición como objeto del pensamiento. En realidad, no podría hablarse de *un* objeto de pensamiento en este primer sentido, pues pensar es un tipo de relación múltiple, una relación a objetos de pensamiento que no son sino los "elementos" de un hecho. La relación epistémica de familiaridad se exige para que un "elemento" llegue a ser objeto de pensamiento en el primer sentido.

6. Pero ¿el razonamiento anterior no asume la verdad del principio en vez de argumentar en su favor? Además, ¿no se hace equivaler implícitamente en él lo que pueda ser un objeto de pensamiento con lo que sean objetos de familiaridad? Un análisis más cuidadoso revela que no es así. Se ha visto que la teoría del juicio y de la comprensión que Russell admite a partir de 1910, considera que los constituyentes de un pensamiento son los *relata* de un tipo relación cognitiva múltiple. Sólo objetos de familiaridad pueden ser *relata* genuinos: si no fuera así, arguye Russell, el pensamiento carecería de "significación". Por tanto, los términos que son *relata* del pensar son objetos de familiaridad si realmente tenemos ese pensamiento. Aquí reside el quicio del argumento russelliano: "Todas las proposiciones que son inteligibles para nosotros, conciernan o no primariamente a cosas que nos son sólo conocidas por descripción, están compuestas de constituyentes con los que estamos familiarizados, pues un constituyente con el cual no estamos familiarizados no es inteligible para nosotros" (Russell, 1910-11, 128). *Lo que* comprendemos o pensamos depende de los objetos de familiaridad que son *relata* genuinos de la relación "relacionante" del pensar, comprender, suponer o juzgar.

Esto, sin embargo, no despeja los interrogantes sobre cuál es la razón para suponer que el objeto *sobre* el que es una descripción definida no es realmente un constituyente de la proposición. He aquí el criterio russelliano: para comprender el significado de una expresión sólo se requiere la comprensión de sus palabras y no la familiaridad con el objeto sobre el cual trata la expresión (Russell, 1918, 193-95). Es, por tanto, necesario que Russell ofrezca una razón independiente de *PF* para defender que la referencia de un nombre o una descripción definida no entra necesariamente como constituyente de la proposición. La razón no puede ser que no es un objeto de

familiaridad y que sólo los objetos de familiaridad sean constituyentes. Aunque Russell parece poner el énfasis en estas razones en varios pasajes de su obra (Russell, 1910-11, 117, 118-119), también señala otras posibles razones: primero, porque es posible conocer una proposición sobre algo sin estar familiarizados con ese algo; segundo, porque es posible comprender proposiciones que incluyen descripciones definidas que no tienen referencia (como es el caso de los enunciados existenciales singulares negativos) (Russell, 1910-11). De nuevo, la primera razón parece circular, pero la segunda es atractiva. Y, como se ha visto, su atractivo corre paralelo al rechazo de no-existentes dentro de su metafísica. Como resalta aún en “The Philosophy of Logical Atomism”, “no puedes tener un constituyente de una proposición que no es nada en absoluto” (Russell, 1918, 242).

7. Lo que se sugiere en los anteriores razonamientos es que *PF* no es un principio que pone en juego esencialmente los objetos de pensamiento como referencias de un pensamiento singular sino los objetos de pensamiento en el primer sentido distinguido anteriormente, como *los términos en que se piensa sobre “algo”*. Pero ¿no choca esto con una lectura natural de las razones que el mismo Russell esgrime a favor de *PF* en *The Problems of Philosophy*?

“Es apenas concebible⁵ que podamos hacer un juicio o sostener una suposición sin saber qué es sobre que estamos juzgando o suponiendo. Debemos atribuir *algún* significado a las palabras que usamos, si vamos a hablar de manera significativa y no emitir mero ruido. El significado que atribuimos a nuestras palabras debe ser algo con lo que tenemos familiaridad” (Russell, 1912, 58)

Parece extraño considerar estas observaciones como *razones* a favor de *PF*. No lo es el hecho de que algunos autores hayan sustituido el principio russelliano de familiaridad por esta razón “intuitiva”, cuya formulación está igualmente repleta de ambigüedades. La “razón” ha sido leída como un *principio de discriminación (PD)* que afirmarí­a que es imposible tener un pensamiento sobre algo sin *saber sobre qué se piensa* (Evans, 1982; McDowell, 1969-70). Es obvio que *PD* no es equivalente a *PF*, pero tampoco el que proporcione un modo de derivar *PF*. ¿Cuál es la relación más plausible entre ambos principios?

Russell plantea el problema al inicio de su artículo “Knowledge by acquaintance and knowledge by description” en términos equívocos: hay que investigar “qué es lo que conocemos en casos en que conocemos proposiciones sobre “el tal-y-cual” sin conocer quién o qué es el tal-y-cual” (Russell, 1910-11, 106). Sea una proposición que conocemos sobre el tal-y-cual, que es *F*, sin saber *qué* es. En cierto modo, la pregunta no deja de ser curiosa, pues parece que podemos afirmar trivialmente que sabemos que el tal-y-cual es *F*. ¿Qué querría decir aquí que no sabemos *qué* es el tal-y-cual?

⁵ En la versión de 1910-11, lo expresa de diferente modo: “parece apenas posible creer...”

Es la aparente ambigüedad en la expresión *saber qué* (en cláusula interrogativa) la que parece generar las confusiones.

Es bien sabido que Austin identificó una doble forma gramatical en este tipo de construcciones: en unos casos, el "qué" ha de ser tomado como un interrogativo; pero en otros, esconde una forma de cláusula de relativo. Así, la cuestión russelliana podría usar el "qué" en dos sentidos diferentes, cuya diferenciación aclararía algo más la búsqueda: ¿*qué* sé cuando sé que el tal-y-cual es F sin conocer *esto que* es el tal-y-cual? Nótese igualmente que lo que se encuentra en juego es un uso escindido de términos epistémicos y que el castellano recoge en una diferenciación entre saber y conocer (presente también en otros idiomas): presupone una distinción entre una interpretación proposicional del saber y una lectura de objeto directo. He aquí entonces, de nuevo, el problema: en casos en que usamos "referencialmente" una expresión como el tal-y-cual y no tenemos un *conocimiento directo* del objeto al que refiere la descripción definida, ¿qué proposiciones sabemos *sobre* el tal-y-cual? Sabemos que el tal-y-cual es F, pero ¿dice la proposición algo que sabemos *sobre* el objeto al que se refiere la expresión el tal-y-cual?

Ahora, parece obvio que comprendemos la expresión "el tal-y-cual es F" aunque no conozcamos (en el sentido directo, de nuevo) el objeto al que refiere la expresión. Por tanto, para Russell, hay algo que sabemos al entender la expresión para que sea significativa. La razón intuitiva que esgrime apela de nuevo a "saber qué es sobre que suponemos o pensamos". Evidentemente, no conocemos *esto que* es el tal-y-cual en la comprensión de la proposición, porque si fuera así, entonces simplemente mediante la comprensión de la proposición de que el tal-y-cual es F, conoceríamos el tal-y-cual mismo, el objeto (sería un nombre en sentido estricto).

Un diagnóstico del problema pretende desenterrar una confusión en el razonamiento russelliano procedente de las dificultades para distinguir las dos relaciones epistémicas de *saber que* y *conocer*: "It looks as though Russell mistakes phrases like "what it is that we are judging or supposing about", as they occur in the latter principle [PD], for expressions designating the appropriate object. This would account for his taking the plausible "*savoir*" principle to be the same as his less plausible "*connaître*" principle" (McDowell, 1998, p. 165). ¿Está aquí el problema russelliano? ¿Confunde ambos principios? ¿Trata transparentemente la relación de *aboutness*?

Una cierta lectura de Russell permite comprobar en estas críticas una infidelidad al espíritu de su argumentación. Quizá el problema y la confusión no residan exactamente en la construcción *saber qué* sino en la expresión total "saber qué es *sobre* que se piensa". Aceptemos que al comprender "lo que suponemos" es que el tal-y-cual es F: una posible lectura apunta al *objeto sobre* el cual es la proposición; otra lectura insiste en *qué* se supone sobre el objeto. La distinción no sería aplicable en el caso de proposiciones que incluyen nombres propios en sentido lógico; y no lo sería porque, dada la

naturaleza de los nombres propios lógicos, sólo están implicados un ruido y el objeto que designa: sólo está presente un *término* que designa de manera directa e inmediata (dada la relación epistémica de familiaridad que está implicada en la comprensión del nombre) un objeto. El *qué* y el *sobre qué* están dados al mismo tiempo.

El error procedería de hacer equivaler todo pensamiento singular a este modelo. Russell no excluye la posibilidad de tener pensamientos singulares sobre objetos con los que no tenemos una relación de inmediatez epistémica: una forma de conocimiento *de cosas mediante* descripción⁶. Son estos casos en los que la distinción se hace visible, incluso cuando están en juego nombres propios ordinarios en los que el pensamiento se debería expresar explícitamente también en términos de una descripción. Son casos de *referencia mediada* bajo los términos de la descripción⁷. PF afirma que estos términos bajo los cuales pensamos sobre algo han de ser objetos de conocimiento inmediato para asegurar la “significatividad” del pensamiento. O lo que es lo mismo: no es posible un sistema de referencias exclusivamente mediado; es necesario apoyar toda referencia mediada en una referencia inmediata, en una relación epistémica *A*⁸.

8. Algunas declaraciones russellianas, sin despejar netamente el misterio, no dejan de ser un soporte para una lectura de este tipo:

“[I]f we are to obtain a description which we know to be applicable, we shall be compelled, at some point, to bring in a reference to a particular with which we are acquainted. Such reference is involved in any mention of past, present, and future (as opposed to definite dates), or of here and there, or of what others have told us. Thus it would seem that, in some way or other, a description known to be applicable to a particular must involve some reference to a particular with which we are acquainted, if our knowledge about the thing described is not to be merely what follows logically from the description” (Russell, 1910-11, 115).

Se sugiere que puede haber dos formas de descripciones definidas (Wettstein, 1990): aquellas de las que se sabe que se aplican a un particular y aquellas que no. Estas últimas no nos dan a conocer más que verdades que se derivan de los términos de la descripción y la proposición que expresan “consta totalmente de conceptos” (Russell, 1910-11, 116), con los cuales el sujeto tiene una relación *A*, por supuesto. En las primeras, está implicada, en algún nivel, una referencia a un particular, una relación directa de familiaridad

⁶ Ya en el artículo de 1905: “The distinction between *acquaintance* and *knowledge about* is the distinction between the things we have presentations of, and the things we only reach by means of denoting phrases” (p. 41).

⁷ Sería inexacto decir que es conceptualmente mediada, pues podría llevar a confusiones: no todos los términos han de ser conceptos; en unos casos, Russell exige un particular.

⁸ Esta lectura abre líneas de interpretación neofregeanas de Russell.

con ese particular: el pensamiento expresado en el uso de tales descripciones (y también y especialmente de ciertos nombres propios ordinarios) incluye un complejo de conceptos y particulares con los cuales el sujeto está familiarizado que logran especificar a quién o qué se aplica la descripción. Lo que sugieren las breves y misteriosas notas de Russell es que el principio de discriminación se satisface al menos de dos formas: o mediante un contacto epistémico directo con el particular en cuestión (nombre propio lógico) o mediante una relación epistémica directa con un complejo de universales y particulares que se aplican al objeto que satisface el predicado que se le atribuye. En un nivel u otro de análisis, es necesaria la presencia de una relación epistémica de familiaridad para asegurar el anclaje de la referencia a particulares o, lo que es lo mismo, la satisfacción del *PD* requiere *PF*.

9. Pero Russell requiere una condición más, que construye a partir de los rasgos genuinamente epistémicos de la relación de familiaridad. Si los términos bajo los cuales pensamos sobre un objeto, bien sea en proposiciones estrictamente singulares que incluyen un nombre propio lógico o bien en proposiciones que implican descripciones definidas de cierto tipo, han de ser *suficientes* para cumplir el requisito impuesto por *PD*, entonces —piensa Russell— los constituyentes del pensamiento han de ser dados para el sujeto de manera transparente e inmediata, sin generar errores. El conocimiento inmediato y no-inferencial que proporciona la familiaridad no puede estar sujeto a duda, bajo amenaza de vaciar el “contenido” del pensamiento. El pensar no puede funcionar en el vacío: sus objetos necesariamente existen, y en un sentido el pensar *conciérne* a esos objetos sin ser esencialmente *sobre* esos objetos.

El núcleo de *PF* como restricción epistémica del “contenido” de cualquier relación cognitiva *R* reside en la asunción cartesiana (Evans, 1982, McDowell, 1986) de que el conocimiento de las propias experiencias tiene la propiedad epistémica de *certeza*. Russell parece requerir que para que un sujeto tenga un pensamiento no pueda estar equivocado sobre los términos en los cuales está pensando al pensar sobre algo: no está abierta al sujeto la posibilidad de identificar erróneamente los objetos del pensamiento. La familiaridad asegura esta característica, puesto que la existencia o inexistencia de los objetos de la relación *A* no está abierta a la duda. Se podría decir que la relación epistémica *A* con un objeto *x* hace necesariamente verdadero “*x* existe”.

Pensar *sobre x* en términos de una descripción definida que sabemos que es aplicable implica, según *PF*, que los términos bajo los cuales se da ese pensamiento sean *objetos* del mismo, pero no que tengamos un pensamiento *sobre* cada constituyente. Sin duda, las expresiones russellianas (to know what we are thinking or supposing about) no son muy felices, puesto que la relación de “aboutness” no ha de aplicarse a los constituyentes del pensamiento. Es más, el razonamiento russelliano sugiere que existe una relación de la mente

con los objetos primitiva y básica cuya caracterización en términos de “aboutness” sería inadecuada: la relación de familiaridad. Russell sustituye una concepción proposicional de la mente por una concepción que hace de la familiaridad la relación cognitiva fundamental. Es el rechazo de la metafísica idealista y monista el que apoya este giro.

10. En el fondo de la sugerencia russelliana, se encuentra el rechazo de una metafísica idealista y monista –especialmente en su dimensión epistemológica– que le condujo a la defensa del atomismo lógico. Para los idealistas, el conocimiento no puede presentarse más que en forma de juicios que aseveran una proposición; Russell defiende la posibilidad opuesta de que todo conocimiento no sea más que un contacto entre la mente y los objetos externos, una relación directa y no-proposicional, a entidades independientes. La familiaridad es una relación cognitiva directa y simple a objetos. Al defender que la comprensión y el juicio – y en general toda relación *R*– son dependientes de la relación primitiva de familiaridad, Russell parece seguir las líneas generales de su programa metafísico.

¿Es el *PF* un instrumento más contra la metafísica idealista? Sólo aparentemente. Más bien, podría afirmarse que la forma que adquiere *PF* en la filosofía russelliana depende del rechazo del idealismo. Hay, al menos, dos asuntos separados en *PF*: el primero concierne a la naturaleza del pensamiento y de sus objetos; el segundo se refiere a la naturaleza de la relación epistémica de familiaridad. Sin embargo, y esto es decisivo, las restricciones epistémicas y metafísicas que caracterizan la familiaridad afectan esencialmente al primer asunto. Parece que es una cuestión abierta cuáles son los *objetos* de la relación de familiaridad y el que un idealista pudiera aceptar, en sus términos, *PF*. Definamos *idea* como *lo que* alguien sabe sobre algo: la definición está de acuerdo con la motivación para aceptar sólo objetos de familiaridad como constituyentes de las relaciones cognitivas complejas, ya que proporciona una restricción sobre el *modo* bajo el cual pueden ser pensados los objetos. ¿Hay alguna razón para no admitir las *ideas* como objetos de familiaridad y, por tanto, que éstas sean *objetos genuinos de pensamiento*? En *The Principles of Philosophy*, Russell presenta un simple argumento en contra de la asunción idealista de que “cualquier cosa que podemos aprehender debe estar en nuestras mentes” (Russell, 1912, 42). Su objetivo es rechazar que las *ideas* sean objetos de *familiaridad inmediata*: de ahí procede su insistencia en la *externalidad* de la relación. Tanto idealistas como subjetivistas han confundido el acto de aprehensión que ocurre en la mente y la cosa aprehendida (que no necesariamente está en la mente). Al hablar de “ideas en la mente”, nos referimos a los *actos* mentales de aprehensión y no a las cosas aprehendidas. Las cosas de las que somos conscientes en un acto de aprehensión no tiene por qué ser mental; la mente misma es caracterizada por su poder de trascenderse a sí misma y contactar con cosas diferentes de sí (Russell, 1912, 42). De modo paralelo, es impropio

hablar de objetos mentales (o incluso entidades mentales); sólo tiene sentido hablar de hechos mentales. "Un hecho será llamado "mental" si contiene o familiaridad o alguna relación que presupone la familiaridad como componente. Así cualquier instancia es familiaridad es mental, ya que es un complejo en el cual un sujeto y un objeto están unidos por la relación de familiaridad. El objeto mismo no necesita ser mental" (Russell, 1914, 162-3). Es más, la aceptación de un velo de ideas generaría un regreso al infinito, pues requeriría una explicación adicional de la relación entre la idea y el objeto, lo cual implicaría una nueva idea, etc.

11. Recojamos los hilos de la argumentación russelliana. *PF* enuncia una restricción epistémica sobre los objetos del pensamiento y, con ello, sobre las condiciones de comprensión de un lenguaje. Esta requiere un conocimiento del significado de las palabras, pues de otro modo serían simplemente ruido. Russell acepta una teoría realista del significado, según la cual los significados son las entidades por las cuales *están* las palabras. Así, la comprensión de un lenguaje exige *saber a qué* entidades refieren. En el caso de los nombres propios en sentido estricto (nombres propios lógicos según la terminología de Russell), se conoce *directamente* a qué entidad refieren: pues ser un nombre propio consiste simplemente en una asociación convencional entre un ruido y un objeto. La comprensión de cualquier expresión compleja depende de la comprensión de sus palabras constituyentes. En expresiones complejas aparentemente referenciales, el conocimiento *de* la entidad a que refieren deriva de una descripción asociada que aplica a un particular propiedades y conocimiento de verdades. Por sí mismo, un conocimiento descriptivo no proporcionaría una comprensión del significado de los símbolos complejos de un lenguaje. Es necesario comprender las palabras que componen la descripción; su significado debe ser algo con lo cual el sujeto debe estar familiarizado. Es su rechazo del idealismo y del monismo el que justifica este paso: no todo conocimiento puede ser de verdades. Dado que la comprensión de un complejo implica un tipo de relación relacionante a los constituyentes *reales* que determinan el modo en que pensamos sobre un objeto, esto es, *qué* es sobre lo que pensamos, y dado que debemos conocerlo y no por medio de otra descripción, es obvio que los componentes deben ser conocidos mediante una relación epistémica más primitiva que Russell denomina familiaridad. Conocer qué es aquello sobre lo que entendemos es estar familiarizado con los términos mediante los cuales nos *acercamos* al objeto del que pensamos: "la descripción es meramente el medio que empleamos para acercarnos lo más posible a la denotación" (Russell, 1910-11, 127). Una vez que sabemos qué términos son los constituyentes de nuestro pensar sobre un objeto, el enunciado que comprendemos no puede significar algo diferente de lo que parecía significar. Podremos aún estar equivocados sobre el objeto que satisface el pensamiento que expresamos, pero no sobre lo que entendemos.

REFERENCES

- Ackerman, F. (1987), "An argument for a modified Russellian Principle of Acquaintance", in J. Tomberlin (ed.), *Philosophical Perspectives 1: Metaphysics*, Atascadero, CA, Ridgeview Publishing Co., 501-512.
- Donnellan, K. (1990), "Genuine names and knowledge by acquaintance", *Dialectica* 44, 99-112.
- Evans, G. (1982), *The Varieties of Reference*, Oxford University Press, Oxford.
- Hintikka, J. (1972), "Knowledge by acquaintance - Individuation by acquaintance", in D. F. Pears (ed.), *Bertrand Russell. A Collection of Critical Essays*, Anchor Books, New York, 52-79.
- Hylton, P. (1990), *Russell, Idealism, and the Emergence of Analytic Philosophy*, Oxford University Press.
- McDowell, J. (1969-70), "Identity Mistakes: Plato and the Logical Atomists", *Proceedings of the Aristotelian Society*, 70, 181-195.
- McDowell, J. (1986), "Singular Thought and the Extent of Inner Space", en P. Pettit y J. McDowell (eds.), *Subject, Thought and Context*, Clarendon Press, Oxford, 137-168.
- Pears, D. F. (1981), "The function of acquaintance in Russell's philosophy", *Synthese*, 46, 159-166.
- Pears, D. (1967), *Bertrand Russell and the British Tradition in Philosophy*, Random House, New York.
- Prior, A. N. (1971), *Objects of Thought*, Oxford University Press, Oxford.
- Russell, B. (1903), *Principles of Mathematics*, Allen and Unwin, London (second ed., 1937).
- Russell, B. (1905/1955), "On denoting", en *Logic and Knowledge*, Routledge, London-New York.
- Russell, B. (1910), "On the nature of truth and falsehood", en *Philosophical Essays*, Longman, Green, and Co., London, 170-185.
- Russell, B. (1910-11), "Knowledge by acquaintance and knowledge by description", *Proceedings of the Aristotelian Society*, 106-129.
- Russell, B. (1912/1959), *The Problems of Philosophy*, Oxford University Press, Oxford.
- Russell, B. (1913/1984), *Theory of Knowledge*, en *Collected papers of Bertrand Russell*, ed. by E. R. Eames and K. Blackwell (ed.), Vol. 7, Allen and Unwin, London.
- Russell, B. (1914/1955), "On the nature of acquaintance", en *Logic and Knowledge*, Routledge, London-New York.
- Russell, B. (1918/1955), "The Philosophy of Logical Atomism", in *Logic and Knowledge*, en *Logic and Knowledge*, Routledge, London-New York.
- Sainsbury, M. (1979), *Russell*, Routledge & Kegan Paul, London.
- Sainsbury, M. (1986), "Russell on acquaintance", in G. Vesey (ed.), *Philosophers Ancient and Modern*, Cambridge, 219-244.
- Wettstein, H. (1990), "Frege-Russell Semantics?", *Dialectica*, 44, 113-135.

Jesús Vega Encabo
 Universidad de Salamanca
 e-mail: jvega@gugu.usal.es

Philosophy of Language

Filosofía del Lenguaje





EL HOLISMO COMO MEREOLÓGÍA*

ANTONIO BLANCO SALGUEIRO
Universidad Complutense de Madrid

ABSTRACT: In the recent literature on semantic holism (Fodor & Lepore, 1992; Devitt, 1996) it is common to assume generic notions of holism that ignore the basic intuition that a holistic property is a property possessed by virtue of belonging to a whole. In this work I suggest a definition of "mereological dependence" that will be used to define the three basic possible doctrines concerning the issue of holism: atomism, molecularism and holism. Then, I apply these distinctions to the field of semantics.

KEYWORDS: Holism, mereology, ontology, semantics, metaphysical dependence, Fodor & Lepore, M. Devitt

1. Introducción

Este trabajo parte de una convicción general: la de que es un buen camino, a la hora de explorar determinadas cuestiones que pertenecen al dominio de la filosofía del lenguaje y que tienen un claro aspecto ontológico, el explicitar la base ontológica que se está asumiendo tanto a la hora de exponer los problemas como a la hora de proponer soluciones a los mismos. Creo que de este modo se pueden tratar mejor algunos problemas sobre los que la literatura filosófica es inmensa, pero respecto a los cuales es a menudo difícil ponerse de acuerdo acerca de las opciones que son conceptualmente posibles. Me refiero a cuestiones que tienen que ver con la naturaleza del significado como son el problema de la naturalización del significado, el del individualismo o el que aquí me ocupará: el problema del holismo semántico.

Partir de la ontología no significa construir esquemas ontológicos desde el sillón para luego aplicarlos a problemas específicos. Mi postura hacia los esquemas ontológicos coincide con la de Jaegwon Kim cuando señala:

"Me inclino (...) a pensar que los esquemas ontológicos son en gran medida optativos, y que las principales consideraciones que

* Este trabajo ha sido realizado con la ayuda de una Beca Postdoctoral de la Fundación Caja Madrid, y dentro del proyecto de investigación financiado por la D.G.I.C.Y.T. con referencia: PB95-0125-C06-02.

deben gobernar la elección de una ontología son las de utilidad, simplicidad, elegancia, y cosas similares.”

(Kim, 1993)

Según esto, un esquema ontológico se ve reivindicado, por ejemplo, cuando sirve para arrojar luz sobre un problema semántico, porque permite plantearlo de modo que las opciones conceptualmente posibles resultan transparentes. Por supuesto, sería ingenuamente optimista pensar que la aplicación de un esquema ontológico a un problema semántico lleva incorporada una solución al mismo. El planteamiento claro de un problema y de las opciones que existen con respecto a él no es sino una precondition de la posibilidad de ofrecer propuestas filosóficas adecuadas.

Mi proyecto no es completamente original. Dos importantes intentos, relativamente recientes, por aclarar la cuestión del holismo semántico parten de una caracterización general, en términos ontológicos, de la cuestión del holismo, para luego buscar una aplicación al caso del holismo semántico. Esto es, se toman en serio el hecho aparentemente evidente de que “semántico” cualifica a “holismo” en la expresión “holismo semántico”, con lo cual difícilmente se podrá avanzar en la clarificación del problema semántico si no se tiene claro el problema genérico del holismo. Me estoy refiriendo a los libros de J. Fodor y E. Lepore (1992): *Holism. A Shopper's Guide* y de M. Devitt (1996): *Coming to Our Senses. A Naturalistic Program for Semantic Localism*.

Hablaré de “la cuestión del holismo” por simplicidad, pero debería quizás decir “la cuestión del holismo-molecularismo-atomismo” por ser estas tres las principales doctrinas que se refieren a la dependencia (en el caso de la semántica) del significado de una expresión con respecto al sistema lingüístico al que pertenece. Los nombres largos son engorrosos y, además, en este caso se corre el riesgo de que existan respuestas posibles no recogidas o mal recogidas en las tres que corresponden al nombre más largo. Del mismo modo, cuando hable de la aplicación de doctrinas metafísicas sobre el holismo a la cuestión del holismo semántico me referiré normalmente a doctrinas metafísicas sobre el holismo, el molecularismo y el atomismo a las cuestiones del holismo semántico, el molecularismo semántico y el atomismo semántico.

Una vez hecho el planteamiento general aclararé la motivación más específica del presente trabajo. Me parece no sólo intuitivamente obvio, sino también ampliamente justificado por los escritos de algunos importantes filósofos holistas, que la doctrina del holismo semántico es una aplicación al campo de la semántica de determinadas tesis mereológicas, esto es, de tesis de dependencia entre partes y todos, en particular en la dirección parte-todo: las partes (mejor, determinadas propiedades de las partes) dependen del todo en el que están contenidas. Así, por poner un solo ejemplo donde aparece con

claridad este planteamiento mereologicista de la cuestión semántica, podemos citar a Gilbert Harman:

“Cualquier versión del holismo del significado dice que existen casos en los cuales las partes tienen significado debido a su lugar en un todo mayor” (“Meaning Holism Defended”, en Fodor y Lepore (eds.), 1993, 163).

Sin embargo, aparte de esta clara orientación mereologicista no es habitual encontrar, entre los holistas semánticos, un camino como el aquí propuesto desde una tesis ontológica general hasta su aplicación semántica. No entraré a discutir el porqué de este desinterés, aunque lo primero que viene a la mente es que es común entre los holistas el considerar que la semántica o la epistemología son prioritarias con respecto a la ontología y que, por tanto, es un error comenzar por esta última.

El caso es que los dos intentos de los que he hablado, que sí siguen el camino por el que me gustaría transitar, coinciden en su antiholismo semántico. Fodor es un atomista semántico y Devitt es un localista con tendencia al molecularismo (el localismo es una tesis que afirma que los significados son o bien molecularistas o bien atomistas, esto es, que no son holistas).

Ambos intentos tienen en común lo siguiente:

- i. Comienzan sentando una base metafísica general que luego aplican a la cuestión del holismo semántico. El holismo semántico, como el molecularismo y el atomismo semánticos, se consideran versiones semánticas del holismo, molecularismo y atomismo entendidos genéricamente como formas de dependencia / independencia ontológica, formas que pueden buscarse en múltiples campos diferentes de la semántica.
- ii. Defienden posturas anti-holistas con respecto al significado.
- iii. Por último, lo que considero más importante para los objetivos de este trabajo, utilizan bases metafísicas en las que desaparece o queda muy oculto el hecho intuitivo antes citado de que el holismo semántico es una cuestión de dependencia de propiedades (en este caso, semánticas) de las partes (esto es, de expresiones lingüísticas, representaciones mentales o creencias) con respecto al todo al cual esas partes pertenecen (un lenguaje, un sistema representacional, o un sistema de creencias o de estados mentales en general).

Mis objetivos en este trabajo serán los siguientes:

- a) Caracterizar las bases metafísicas que respaldan los planteamientos de la cuestión del holismo semántico en Fodor & Lepore (1992) y en Devitt (1996), así como su aplicación a la semántica. La caracterización que haré será informal, por brevedad y por no hacerme pesado con definiciones a veces engorrosas, pero es posible ofrecer una reconstrucción más formal, detallada y

sistemática.¹ Defenderé que ninguno de esos dos intentos acierta a plantear el problema en sus justos términos.

b) Proponer una base metafísica que nos permita reconstruir la polémica acerca del holismo en términos explícitamente mereológicos. La construcción de esta base plantea, a mi entender, dos problemas principales para los cuales considero que es posible ofrecer soluciones adecuadas. El primero consiste en construir una base ontológica que permita distinguir entre dependencia moderada y dependencia extrema de las propiedades de una parte con respecto al todo del que forma parte (esto es, que permita distinguir entre holismo y molecularismo). El segundo consiste en construir una definición de “dependencia mereológica” que no sea circular o trivial.

2. Fodor & Lepore: holismo como extrema anatomicidad

Fodor y Lepore parten de la noción de “anatomicidad”, que definen de la siguiente forma: “una propiedad es anatómica sólo en el caso de que si algo la posee, entonces al menos otra cosa la posee” (Fodor y Lepore, 1992, p. 1). El ejemplo de propiedad anatómica que proponen es el de *sibling* (palabra sin equivalente primitivo en castellano, podríamos traducirla por *hermano o hermana*, o por *hermano* en su sentido genérico). Necesariamente, si algo es un hermano (en este sentido genérico) entonces existe otra cosa que también es un hermano.

Definen entonces los autores las tres doctrinas metafísicas principales del siguiente modo (los ejemplos son de Fodor y Lepore):

* Una propiedad es **holista** syss es muy o extremadamente anatómica. Ej: *número natural*, definida por la propiedad de tener un sucesor.

** Una propiedad es **molecular** syss es moderadamente anatómica. Ej: *sibling*.

*** Una propiedad es **atómica** syss no es anatómica. Ej: *ser una roca*.

Es fácil ver que estas definiciones no son explícitamente mereológicas. Una propiedad podría, en principio, ser extremadamente anatómica sin que las distintas entidades que instancian tal propiedad perteneciesen a un mismo todo abarcador.

Las doctrinas semánticas correspondientes no son difíciles de definir. Para el caso más simple, tendríamos que las propiedades semánticas (genéricas) son holistas syss se da el caso de que si algo (una expresión) las instancia, entonces han de existir muchas otras cosas (expresiones) que también las instancien. Serían moleculares syss hubiesen de existir algunas otras expresiones significativas y atómicas si fuese posible que una expresión significase en ausencia de otras expresiones. Las doctrinas de Fodor y Lepore se formulan para propiedades semánticas genéricas, no para propiedades

¹ Realizo esa exposición más detallada y formal en Blanco Salgueiro (próxima publicación).

semánticas específicas. Sería absurdo afirmar el holismo con respecto a una propiedad semántica específica como *significar que la nieve es blanca*, aunque no con respecto a una propiedad genérica como *significar*. Esto es, es peregrino afirmar que si algo significa que la nieve es blanca otras muchas cosas han de significar que la nieve es blanca, aunque no lo es afirmar que si una cosa significa entonces muchas otras cosas significan. De todos modos, es éste un rasgo extraño del planteamiento de Fodor y Lepore. Normalmente nos interesa hablar de propiedades semánticas específicas, del hecho de que éstas dependan o no de que las expresiones que las instancian pertenezcan a sistemas simbólicos. La imposibilidad de plantear la cuestión en esos términos más intuitivos es algo que se deriva del planteamiento metafísico, del hecho de haber elegido la noción de “anatomicidad” como la pertinente para el planteamiento de la cuestión del holismo semántico. Tenemos aquí una prueba de que la elección de una base metafísica impone restricciones importantes al planteamiento de los problemas semánticos.

3. Devitt: holismo como instanciación de múltiples propiedades

El planteamiento que hace Devitt es claramente distinto al de Fodor y Lepore. Para este autor, el *quid* de la cuestión reside en el número de otras propiedades que se exige que un individuo instancie para instanciar una determinada propiedad. Así, la propiedad de *ser un capitalista* exige, de aquel objeto que la instancie, que también intancie otras propiedades diferentes, por ejemplo la propiedad *poseer medios de producción*. Así, tendríamos definiciones generales como las siguientes:

* Una propiedad es **holista** syss su instanciación por parte de *x* requiere la instanciación por parte de *x* de un número elevado de otras propiedades. Ej: *poseer un ejemplar de cada una de las monedas de curso legal que hayan existido hasta 1999*.

** Una propiedad es **molecularista** syss su instanciación por parte de *x* requiere la instanciación por parte de *x* de un número moderado de otras propiedades. Ej: *capitalista* (y, según Devitt, la mayoría de las propiedades que interesan a las ciencias).

*** Una propiedad es **atómica** syss su instanciación por parte de *x* no requiere la instanciación por parte de *x* de otras propiedades. Ej: *tener la carga z*.

Claramente, las taxonomías de Fodor & Lepore y de Devitt no son equivalentes, ya que no agrupan a las mismas propiedades dentro de las mismas categorías. Por poner un solo ejemplo, aunque parece claro que *capitalista* es atómica en el sentido de Fodor & Lepore, puesto que podría existir un único capitalista, esa misma propiedad es molecularista en el sentido de Devitt si, como afirma ese autor, para ser un capitalista es necesario poseer

algunas (pero no muchas) otras propiedades (por ejemplo, *poseer medios de producción*).

La aplicación a la semántica se realiza definiendo en primer lugar la noción de *propiedad inferencial* y admitiendo que cualquier expresión (de un lenguaje natural como el castellano) posee en general un número elevadísimo de propiedades inferenciales. Las propiedades inferenciales se derivarían de las diferentes conexiones epistémicas entre palabras (o entre los conceptos que les corresponden). Habría una tal conexión entre “gato” y “animal”, pero quizás también (para un determinado hablante), una conexión entre “gato” y “animal adorado por la abuela María”. Una propiedad semántica será holista, molecularista o atomista según que el número de propiedades inferenciales constitutivas (de los que depende) la instanciación de esa propiedad sea elevado, moderado o nulo.

Como se puede apreciar, tampoco existe nada en este planteamiento que sea explícitamente mereológico. Obviamente, si existen propiedades inferenciales, existirán expresiones diferentes de aquella de la que se predica la propiedad semántica, pero ni siquiera se requiere explícitamente que pertenezcan a un mismo todo. Esto se deriva de que no se requiere algo parecido en la definición ontológica general.

Por tanto, tampoco el planteamiento de Devitt se recoge la intuición de que el holismo es una cuestión de dependencias entre partes y todos. Ahora bien, las doctrinas que se pueden reconstruir a partir de las definiciones de Devitt sí se aplican a propiedades semánticas específicas como *significar que la nieve es blanca*. Esa propiedad es atomista si el hecho de que una expresión signifique que la nieve es blanca es independiente de sus propiedades inferenciales, es molecularista si se precisan algunos lazos inferenciales, y es holista si se necesita un número elevado de tales lazos.

4. Holismo como extrema determinación mereológica

Por último, ofreceré mi propia base metafísica para el planteamiento de la cuestión del holismo. Como indiqué en la introducción, existen a mi entender dos dificultades principales a la hora de plantear en términos mereológicos la cuestión del holismo semántico. La primera de ellas tiene una solución relativamente fácil. La segunda no tiene tan fácil solución y puede explicar los rodeos de Fodor y Lepore y de Devitt a la hora de presentar una base metafísica sin referencia explícita a todos y partes:

1. Construir una base metafísica que permita distinguir entre molecularismo y holismo. El problema reside en que así como el atomismo se puede definir como independencia de la instanciación de una propiedad con respecto a la pertenencia del instanciador a un todo, tanto el molecularismo como el holismo son formas de dependencia con respecto al todo. Ahora bien,

¿Cómo distinguir entre una dependencia “extrema” y una dependencia “moderada” con respecto a un todo?

2. Existe el problema más general de definir no trivialmente las condiciones metafísicas en las que algo forma parte de un todo. Algunos autores, de hecho, han defendido que es imposible salirse del círculo mereológico: que no es posible definir las condiciones en que varios *xs* componen un *y* sin apelar a otros conceptos mereológicos (como todo, parte, etc.)².

Empecemos por el primer problema. Para ponerlo de relieve podemos partir de una definición de dependencia mereológica que es inadecuada:

(M. 1) Una propiedad *P* **depende mereológicamente**₁ de una propiedad *Q*
=def.

i. El hecho de que un *x* instancie *P* depende metafísicamente de que exista un *y* tal que *y* instancia *Q*.

ii. *x* es una parte propia de *y*.

Ejemplo (de Husserl): *color* depende mereológicamente de *cosa coloreada*.

Ejemplo de independencia: *grano de arena* no depende mereológicamente de *montón de arena*.

Parece evidente que un grano de arena, aun en el caso de que cumpla la condición ii. y forme parte de un montón de arena, seguiría siendo un grano de arena aunque no perteneciese a ese montón (o a cualquier otro), con lo cual no se cumple la condición i. En cambio esto no sucede en el caso de *color* y *cosa coloreada*. Según Husserl, el color y la extensión son dos momentos mutuamente dependientes de un todo mayor, de modo que el color no puede existir sin una extensión (ni viceversa).³

Basándonos en esa definición podemos entonces distinguir entre propiedades que dependen de la existencia de un todo englobante de aquellas otras que son independientes de tal pertenencia. Así, podríamos definir el “atomismo” en términos de independencia mereológica₁ absoluta: una propiedad atomista no depende mereológicamente₁ de ninguna otra propiedad. Pero esto es insuficiente, ya que también necesitamos distinguir entre propiedades molecularistas y holistas. No sólo necesitamos una distinción entre propiedades sensibles y propiedades insensibles a la pertenencia a un todo, sino que tenemos que distinguir además entre propiedades que son *extremadamente* sensibles a los cambios en el todo englobante de otras que son *moderadamente* sensibles a tales cambios. A fin de responder a esta necesidad defino una nueva noción de *mereologicidad*:

² Ver van Inwagen (1990) o Markosian (1998).

³ Ver Husserl (1982).

- (M. 2) La propiedad P **depende mereológicamente₂** del par de propiedades $\langle Q, T \rangle$ si y sólo si para todo x , el hecho de que x instancie P depende metafísicamente de que existan un y y un z tales que:
- El hecho de que x instancie P depende metafísicamente de que exista un y ($y \neq x$ y y distinto de una parte propia de x), tal que y instancia Q .
 - x e y son partes propias de z .
 - z instancia T .

Ej (putativo): *corazón* depende mereológicamente₂ de $\langle \text{sangre}, \text{organismo} \rangle$

Ej (husserliano): *color* depende mereológicamente₂ de $\langle \text{extensión}, \text{cosa coloreada} \rangle$

Armados con la definición M. 2 podemos distinguir entre molecularismo y holismo mereológicos, según la cantidad de pares $\langle Q_i, T_j \rangle$ de los cuales dependa una propiedad:

* La propiedad P es **holista** syss existen muchas propiedades $T, Q_1 \dots Q_n$ tales que P depende mereológicamente₂ de cada par en el conjunto $\{\langle Q_1, T \rangle \dots \langle Q_n, T \rangle\}$

** La propiedad P es **molecular** syss existen sólo unas pocas propiedades $T, Q_1 \dots Q_n$ tales que P depende mereológicamente₂ de cada par en el conjunto $\{\langle Q_1, T \rangle \dots \langle Q_n, T \rangle\}$

*** La propiedad P es **atómica** syss no existe ningún par $\langle Q, T \rangle$, tal que P depende mereológicamente₂ de $\langle Q, T \rangle$.

Así pues, parece que el primero de los problemas que nos habíamos planteado tiene una solución relativamente sencilla. El segundo, a mi entender, también es tratable. Discrepo frente a los que consideran que es imposible una definición no trivial o no circular de “dependencia mereológica”. De hecho, considero que la definición que he dado más arriba no es trivial, a pesar de que en ella se utiliza la noción de “parte propia” para definir la de “dependencia mereológica”. En la literatura sobre composición, uno de los principales problemas es el de distinguir las posturas “universalistas”, las cuales afirman que cualesquiera objetos, por dispares que sean (por ejemplo, mi zapato izquierdo y la ciudad de Granada), constituyen un nuevo objeto consistente en su suma mereológica, de posturas más restrictivas que defienden que sólo algunas cosas se unen para formar todos genuinos. La noción de parte que utilizo en mis definiciones es la noción menos restrictiva, la que aceptaría un universalista. Pero en la condición i. se utiliza una noción ontológicamente más básica, la de “dependencia metafísica”, para establecer las restricciones pertinentes. Es esta condición la que nos permite decir que un montón de arena no es un auténtico todo, mientras que una cosa coloreada sí lo es. Por tanto, es aquí donde no debe haber un concepto explícita o implícitamente mereológico, a riesgo de que nuestra definición de dependencia mereológica sea trivial o circular. Y parece claro que la noción de dependencia metafísica, si

tuviera que ser definida (esto es, en el caso de que no la adoptásemos como primitiva) no incorpora ninguna noción mereológica.

Por otra parte, la noción de dependencia metafísica es utilizada también en los planteamientos de Fodor y Lepore y de Devitt. En un caso, se trata de la dependencia de la instanciación de una propiedad con respecto a su instanciación múltiple; en el otro, se trata de la dependencia de la instanciación de una propiedad por parte de un individuo con respecto a la instanciación de muchas otras propiedades por parte de ese mismo individuo. Se trata de una noción que es probable que aparezca en cualquier planteamiento ontológicamente enraizado de la cuestión del holismo semántico.

5. Conclusiones

Concluyo que una vez superados los dos problemas que he señalado en el apartado 4, es preferible un planteamiento mereologicista de la cuestión del holismo semántico, ya que ello supone una mayor proximidad con respecto a las intuiciones que tenemos acerca de la dependencia o independencia de las propiedades semánticas de los símbolos y, por otra parte, un mayor respeto a los planteamientos semánticos de los filósofos que han defendido posturas holistas con respecto al significado.

BIBLIOGRAFÍA

- Blanco Salgueiro, A. (próxima publicación): "Relacionalidad y atomicidad".
Devitt, M. (1996): *Coming to Our Senses. A Naturalistic Program for Semantic Localism*, Cambridge, Mass.: Cambridge University Press.
Fodor, J. A. y Lepore, E. (1992): *Holism: A Shopper's Guide*, Oxford, Blackwell.
Harman, G. (1993): "Meaning Holism Defended", en Fodor, J. Y Lepore, E. (eds.) (1993): *Holism: A Consumer Update*, Amsterdam, Rodopi, pp. 163-172.
Husserl, E. (1982): *Investigaciones lógicas 2*, Investigación tercera: "Sobre la teoría de los todos y las partes", Madrid, Alianza, pp. 385-433.
Kim, J. (1993): *Supervenience and Mind*, Cambridge, Mass.: Cambridge University Press.
Markosian, N. (1998): "Brutal Composition", *Philosophical Studies*, 92, pp. 211-249.
Van Inwagen, D. (1990): *Material Beings*, Ithaca, Cornell University Press.

Antonio Blanco Salgueiro
Universidad Complutense de Madrid
e-mail: anblasal@usc.es



SIGNIFICADO: PRIVADO, PÚBLICO, COMPARTIDO

CRISTINA CORREDOR
Universidad de Valladolid

ABSTRACT: Aim of this paper is to discuss a recent essay by E. Craig (1997) in which the author argues in favor of a form of meaning privacy, and for the fact that a semantics based in 'epistemically private items' is part of any public language. But two objections seem possible here and both are briefly discussed. In particular, it can be objected that this internalist account of meaning does not distinguish two different perspectives on linguistically mediated interaction: a descriptivist one, in terms of the observable coordination of speakers, and the Kantian perspective of what is *de iure* in any valid linguistic understanding -particularly so when it is not a theoretical discourse, but a practical one (a juridical or a moral discourse) what is in question. Finally, it is argued in favor of a normativist view on meaning and communication, where the presupposition of a shared language is to be seen as a 'quasi' regulative premiss, constitutive of human understanding.

KEYWORDS: Philosophy of language, rule-following, private language, intentionalism, normative, normativist view on meaning, Wittgenstein, E. Craig, C. Wright, E. Husserl

El objetivo de la presente comunicación es discutir un reciente ensayo de Edward Craig¹ en el cual el autor argumenta, en primer lugar, a favor de una forma de privacidad del significado, y, en segundo lugar, defiende que una semántica basada en "elementos epistémicamente privados" (*epistemically private items*) es parte de todo lenguaje público. El contexto teórico presupuesto es el del *argumento del lenguaje privado* de Wittgenstein, y dos importantes desarrollos en teoría del significado que toman un punto de partida en aquél: la interpretación *comunitarista* de S.A. Kripke² de la noción de seguir una regla, y la semántica *antirrealista* de M. Dummett. La discusión

¹ Craig, E.: "Meaning and privacy", en *A Companion to the Philosophy of Language*, B. Hale y C. Wright eds., Blackwell, Oxford, 1997, pp. 127-145.

² Kripke, S.A. (1982): *Wittgenstein on rules and private language*, Oxford, Blackwell.

entraña consecuencias para una de las tesis centrales y más controvertidas de la Filosofía Analítica: la *tesis de la prioridad* propuesta por M. Dummett.

I

El conocidísimo *argumento del lenguaje privado* de Wittgenstein³ puede reconstruirse a partir de dos ideas: en primer lugar, que los enunciados de un "lenguaje privado" para las sensaciones sólo serían plenamente significativos si pudiera establecerse una distinción sustantiva entre lo que le parece correcto decir o pensar al usuario de este lenguaje privado, y lo que es realmente correcto en él; y, en segundo lugar, que esta distinción no es posible. Como ha señalado C. Wright⁴, esta conclusión remite al problema de cuál es la naturaleza del conocimiento de nuestros propios contenidos intencionales; pues, de acuerdo con la crítica de Wittgenstein, una identificación privada correcta de una sensación, por ejemplo, haría entrar en juego la aplicación del símbolo lingüístico que identifica a cualquier sensación que sea del mismo tipo. Pero aquí surge una dificultad, que obliga a distinguir entre dos tipos de conocimiento posible, y a tener en cuenta a ambos. Por una parte, puesto que nuestros estados intencionales son testificables, son susceptibles de expresarse en el medio que representa el lenguaje, ellos son conformes al paradigma de la sensación y de otros fenómenos de la conciencia que pueden considerarse observables por introspección. Pero, por otra parte, también son conformes, constitutivamente, a los modos en que el sujeto los manifiesta, y en esta medida son conformes al paradigma de rasgos psicológicos que, como la modestia o la irritabilidad, hay que considerar disposicionales: esto significa que no dan lugar preeminentemente a testimonios, y que el procedimiento para identificarlos no es la introspección, sino la inferencia a partir de la observación perceptiva externa. En el primer caso, el de la observación introspectiva, el sujeto parece gozar de una posición privilegiada en el acceso a sus estados intencionales -lo que se ha llamado la *autoridad de la primera persona*-; pero, en el segundo caso, este mismo sujeto no está en mejor posición que un observador externo que pueda realizar las mismas inferencias.

Esto ha llevado a constatar que un presupuesto básico que ponemos en juego cuando atendemos a la epistemología de los estados intencionales, el de la autoridad comúnmente atribuida a cualquier sujeto respecto a sus propios

³ Wittgenstein, L.: *Investigaciones Filosóficas*, Barcelona y México, Crítica, 1988, props. 243-315.

⁴ Cf. Wright, C.: "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention", en *Meaning Scepticism*, K. Puhl ed., Berlín, de Gruyter, 1991, pp. 126-147.

estados intencionales, ha de verse como un *principio constitutivo*: algo "que no es una *consecuencia* de la naturaleza de esos estados, ni una relación epistémicamente privilegiada asociada [con ellos], sino que se introduce con carácter primitivo entre las condiciones de identificación de lo que un sujeto cree, espera y pretende"⁵. Es importante tener en cuenta esta caracterización porque el punto de partida en el ensayo de Craig que se va a comentar es exactamente este mismo presupuesto: el de que los seres humanos son capaces de encontrarse en determinados estados intencionales, a los que tienen un acceso privilegiado, y que "son privados, en el sentido técnico de que su naturaleza no la puede conocer más que el sujeto que los experimenta"⁶. Es importante darse cuenta de que para Craig este conocimiento identificativo no presupone la existencia de un lenguaje de símbolos, que permita identificar a los estados intencionales que son del mismo tipo que uno dado. Por el contrario, su propósito es defender una concepción intencionalista del significado; por ello, su discusión se articula en torno a una pregunta que él mismo formula: la de si la naturaleza de nuestros *estados intencionales privados* (en adelante, EIPs) afecta, o puede afectar, a la semántica de una lengua natural empleada para la comunicación interpersonal.

El difícil y fragmentario texto en el que Wittgenstein argumenta en contra de que pueda haber un lenguaje privado ha encontrado reelaboración ulterior en el trabajo de M. Dummett, quien distingue y enuncia varios argumentos. Todos ellos, conjuntamente, caracterizan la propia *semántica antirrealista* de Dummett, posición que consituye de hecho una respuesta negativa a la pregunta formulada por Craig: pues se encaminarían a rechazar la idea de que los EIPs puedan desempeñar alguna función en la semántica de una lengua natural pública. De acuerdo con el razonamiento de Dummett⁷, si el significado de una expresión empleada por un hablante estuviera parcialmente determinada por lo que, por su carácter epistémicamente privado, sólo a él le es accesible, lo que sólo él puede conocer, entonces sería imposible que pudiera comunicarse mediante esa expresión con otros; pues nadie más podría entenderla y no podría ser, por tanto, un elemento de la lengua pública. Si el significado se hace depender de EIPs, entonces el conocimiento de lo significado por un hablante con una expresión es inaccesible por principio y, por tanto, incognoscible para cualquier otro

⁵ Wright (1991), *ibid.*, p. 142.

⁶ Craig (1997), *ibid.*, p. 127.

⁷ Dummett, M. (1973): "The philosophical basis of intuitionistic logic", en *Truth and other enigms*, Londres, Duckworth, 1978, pp. 215-47.

hablante, y el lenguaje o la lengua empleados difícilmente pueden considerarse un medio para el entendimiento mutuo y la comunicación interpersonal.

Craig considera, sin embargo, que esta conclusión está mal fundada. Él comienza discutiendo la posición de Wittgenstein, y defiende una definición de *lenguaje privado* -de las dos que cabe discernir en *Investigaciones*, prop. 243- que hace de éste un fragmento constitutivo de todo lenguaje público. Su razonamiento se apoya de manera explícita en lo que el autor llama la *asunción de Burke*, o hipótesis de Burke, y que no es sino una versión del famoso *argumento analógico*. De acuerdo con él, “suponemos de modo natural que los otros, en circunstancias externas similares en un sentido amplio, tienen estados internos similares en un sentido amplio”⁸. Craig llega finalmente a la conclusión -tras considerar críticamente la interpretación de Kripke y su solución comunitarista escéptica- de que los hablantes asignan significado a fragmentos amplios del lenguaje público asociando las expresiones lingüísticas correspondientes con sus propias vivencias, sensaciones o experiencias en un sentido amplio -con sus EIPs-; y sólo después, por una suerte de transferencia analógica, generalizan y suponen que también los demás tienen experiencias, sensaciones y vivencias análogas cuando emplean las mismas expresiones, es decir, suponen que sus EIPs son las mismas o, al menos -con cautela escéptica-, suficientemente parecidas. Esta hipótesis se mantendría en la medida en que, a partir de ella, la comunicación con los otros, lingüísticamente mediada, se logra establecer y se mantiene con éxito.

Pero importa observar que, en la reconstrucción de Craig, *no hay garantía alguna de una identidad intersubjetiva del significado* -identidad de lo significado por distintos hablantes con las mismas expresiones- y sí, por contra, una imposibilidad de *conocer*, en un sentido enfático, lo significado por los otros hablantes. Esta discusión de Craig resulta interesante porque acierta a hacer perfectamente explícito un presupuesto que estaba presente tanto en la interpretación comunitarista de Kripke (1982) como en la semántica antirrealista de Dummett: el requisito de una noción *fuerte* de conocimiento, en relación con el saber del significado de los hablantes. Lo

⁸ Craig (1997), *ibid.*, p. 131.- El ensayo de Craig que se está comentando es la expresión última de un trabajo detenido y extenso en el tiempo que se ha venido ocupando del mismo tipo de problemas. Como precedentes inmediatos cabe citar los siguientes. Craig (1982): “Meaning, use and privacy”, en *Mind* 91 (1982), pp. 541-64; también Craig (1986): “Privacy and rule-following”, en *Language, mind and logic*, J. Butterfield ed., Cambridge, Cambridge U.P., pp. 169-86.

que tácitamente asumen los detractores de un lenguaje privado es que el entendimiento mutuo sólo es posible si hay un conocimiento completo y correcto de lo que las expresiones empleadas significan *para el otro*. Este presupuesto es precisamente al que está dispuesto a renunciar Craig: los hablantes pueden, sobre la base de una transferencia analógica de los propios estados intencionales al otro, establecer con éxito pautas de interacción, aunque no haya “garantía alguna de mutuo entendimiento; pero ¿por qué no podría haber entendimiento mutuo sin esa garantía?”⁹.

La idea de Craig parece ser la de que el saber del significado de una expresión incluye dos componentes, asimétricos y diferenciados: el componente del propio EIP asociado con la expresión, del que es posible un conocimiento por observación introspectiva, y el componente del EIP que atribuimos a otros cuando emplean la misma expresión, cuyo conocimiento - por seguir con el mismo esquema de Wright- es inferencial. Esta reconstrucción supone, sin embargo, aceptar no sólo que el conocimiento de lo significado en el lenguaje público es siempre y necesariamente parcial -se infiere lo significado por los otros a partir de los “reajustes” que es preciso hacer sobre la primera transferencia analógica, al observar su conducta externa-; supone aceptar, además, que nuestra confianza en que nos entendemos con los otros cae “fuera del alcance de una justificación racional”.

Por eso mismo, la posición esbozada por Craig tiene el efecto inmediato de una renuncia al conocimiento *reflexivo*, o de “segundo orden”, respecto a en qué consiste el propio saber del significado: “No hay razón particular para creer que, para entender correctamente, uno tenga que saber que entiende correctamente”¹⁰. Por el contrario, el requisito de identidad intersubjetiva del significado sí que descansa en la necesidad, al menos lógica o conceptual, de un conocimiento de “segundo orden” respecto a qué significa emplear correctamente la expresión correspondiente -en Dummett, qué cuenta como justificación suficiente de un enunciado teórico o práctico-, para que el significado se constituya en el uso y para que pueda decirse, del lenguaje público, que es un lenguaje *compartido*. Frente a esta posibilidad, que caracterizaría a una concepción pragmatista y *normativa* del significado, la idea de Craig parece ser la de que la transferencia por analogía, y las inferencias que la siguen, se ven corroboradas por el éxito en la interacción, y que esto es algo constatable fácticamente. Ante las posibles objeciones

⁹ *Ibid.*, p. 131.

¹⁰ *Ibid.*

formuladas contra su propuesta, por consiguiente, su defensa consiste en apelar a los hechos.

II

Al formular lo que llama asunción o hipótesis de Burke, Craig recoge una cita de este autor en la que encuentra una formulación más clara y explícita del supuesto en cuestión: “Suponemos, y tenemos que hacerlo así, que, así como la conformación de sus órganos son prácticamente o por entero los mismos en todos los hombres, así también el modo de percibir los objetos externos es el mismo en todos los hombres, o con ligeras diferencias”¹¹. La propuesta de Craig depende de manera esencial de esta hipótesis, que acompaña al conjunto de posiciones que, en teoría del significado, podrían considerarse formas de un *internismo moderado*. Quizá no sea completamente paradójico que esta misma hipótesis aparezca en un autor que podría considerarse, en principio, muy alejado de los presupuestos y del método de la Filosofía Analítica, contexto en el que sí se encuadraría toda la discusión precedente. Y, sin embargo, en la quinta de sus *Meditaciones Cartesinas*, y al intentar reconstruir la intersubjetividad de la comunicación lingüística tomando como punto de partida la intencionalidad de la conciencia individual, E. Husserl tuvo que hacer entrar en juego la misma hipótesis de Burke: la de una transferencia analógica.

Dejando al margen la complejidad filosófica del padre de la Fenomenología, y sin hacer justicia seguramente a la significación global de su obra, sí parece posible fijarse en lo que parece compartir con Craig y con otros autores analíticos: lo que el propio Husserl llamó *solipsismo metodológico*. Desde esta perspectiva, él intentó mostrar cómo los rendimientos de una conciencia constituyente, que actúa monológicamente, son previos y más fundamentales que los significados ligados a un lenguaje público. Pero su reconstrucción, como han puesto de manifiesto sus críticos, adolece de dificultades teóricas esenciales. Lo que interesa ver ahora es que esas mismas objeciones les son aplicables a cualquier concepción del significado que, partiendo del mismo solipsismo metodológico, se apoye de forma análoga en la hipótesis de Burke.

Husserl¹² parte de la premisa de que, con respecto a los otros sujetos, lo que llama el *yo constituyente* -el yo del fenomenólogo, que se ha sometido

¹¹ E. Burke, 1757; cit. en Craig (1991), *ibid.*, n. 3 en p. 144.

¹² E. Husserl, “Cartesianische Meditationen”, en *Husserliana*, vol. 1, La Haya, 1950.- La crítica que se va a recoger fue formulada originalmente por A. Schütz (1957): “Das Problem der transzendentalen Intersubjektivität bei Husserl”, en *Philosophische Rundschau* 5/2

previamente a la disciplina de sucesivas *epojés* que le liberen de todo presupuesto y de todo pre-juicio- constata que todo significado tiene lugar a partir de su propia “vida intencional”, que procede de sus síntesis constitutivas. El problema es cómo tiene lugar, desde el yo constituyente, la constitución de otra subjetividad, y de todo sentido con validez intersubjetiva; para responder a ello, Husserl se propone mostrar cómo es posible derivar la intersubjetividad (objetividad) del mundo a partir de la intencionalidad de una conciencia. La reconstrucción de Husserl imagina una situación contrafáctica, en la que lo que aparece ante el yo constituyente es un “mundo primordial” en el que sólo está contenido lo que le es dado a éste. La otra subjetividad, que en un primer momento sólo aparece como una mera corporalidad, se constituiría en dos etapas. En una primera, es una *aprehensión por analogía* la que permite transferir, a otros cuerpos, las mismas experiencias originarias que el yo constituyente tiene de su propia corporalidad. Husserl habla aquí, precisamente, de una *transferencia por analogía*; y a su resultado, por el cual al yo constituyente le es “co-presente” algo de lo que no tiene experiencia originaria -la interioridad inaccesible del otro-, lo llama *apresentación*, y es en esta etapa donde queda constituido el sentido de otro sujeto. En una segunda etapa, haciendo entrar en juego un nuevo supuesto: el de la *intercambiabilidad de perspectivas* -por analogía explícita con los sistemas de referencia espaciales-, Husserl supone que al yo constituyente le es posible constituir el mundo y la interioridad del otro, al establecer analogías con los suyos propios: es decir, al imaginar que ocupa el “aquí” y el “ahora” del otro, puede transferirle las vivencias que él cree que tendría si ocupara la misma posición. De este modo resultaría, finalmente, lo que Husserl llama un “nosotros transcendental”.

Para que sea posible la transferencia por analogía que Husserl atribuye al yo constituyente, él supone en su reconstrucción que las sucesivas apresentaciones del otro, los modos en que se manifiesta, se caracterizan por una regularidad significativa y por un carácter concordante al que el yo constituyente puede atribuir un sentido. Así, afirma: “La corporalidad vivida extraña de la que se tiene experiencia sólo se manifiesta realmente como corporalidad vivida por la continuidad en su conducta cambiante, pero siempre *concordante*; de tal modo que ésta tiene su lado psíquico, que el

(1952), pp. 81-107. Una revisión posterior se debe a J. Habermas (1970/71): “Phänomenologische Konstitution der Gesellschaft...”, en *Vorstudien und Ergänzungen zur Theorie des komm. Handelns*, Francfort, Suhrkamp, 1989, pp. 35-59. También se ha ocupado de ello W. Stegmüller (1989): *Hauptströmungen der Gegenwartsphilosophie*, vol. 1, cap. 2, Stuttgart.

físico indica al presentarlo (...) La corporalidad vivida se experimenta como sólo aparente cuando, con el cambio [en la conducta], no se preserva la concordancia”¹³.

Es fácil reconocer, en esta reconstrucción detallada, el mismo tipo de proceso que Craig sugería más arriba. Pues en ambos casos entran en juego los dos supuestos que van a ser centrales para la crítica: la asunción de Burke, o movimiento de transferencia analógica, y un segundo supuesto que Husserl ha hecho explícito: el de la intercambiabilidad “geométrica” de las perspectivas del sujeto constituye y su interlocutor. Por ello mismo, las objeciones que se han podido formular a la propuesta de Husserl son aplicables a Craig, y a cualquier concepción guiada por los mismos supuestos. Estas objeciones son, fundamentalmente, dos. En primer lugar, se ha observado que la regularidad en la conducta que Husserl reclama, para que los movimientos o las manifestaciones del otro puedan interpretarse como símbolos con un significado, *presupone precisamente aquello que supuestamente se va a constituir*: porque, si se quiere reconstruir un lenguaje común a partir de los rendimientos de una conciencia solipsista y autónoma, que según la hipótesis del solipsismo metodológico sería previa y más originaria que cualquier lenguaje, no puede suponerse que el yo constituyente dispone ya del concepto de un sistema simbólico, ni de alguna experiencia comunicativa previa que le permita considerar, a las manifestaciones externas puramente físicas, como signos con un significado. Pero esto es exactamente lo que supone Husserl, como también lo supone tácitamente Craig en su explicación: ambos dan por supuesto que el sujeto dispone de un saber previo de algún sistema simbólico -arquetípicamente, una lengua-, y que dispone asimismo de la experiencia de lo que significa emplear signos para comunicarse con ellos, con otro sujeto, sobre algo en el mundo.

La segunda crítica formulada contra Husserl, y aplicable igualmente a Craig, se refiere al supuesto de la intercambiabilidad de perspectivas. Pues se ha hecho notar, con razón, que este supuesto podría ser aceptable -si lo es- en el caso de un discurso teórico, basado únicamente en conceptos empíricos. Pero este supuesto, guiado por el modelo de la percepción sensorial, precisa que se dé una simetría completa entre las perspectivas del sujeto y su interlocutor: y difícilmente esta posibilidad de intercambio de perspectivas puede extrapolarse al mundo social. Pues esto sólo sería plausible si, de nuevo previamente, se hubiera establecido ya una reciprocidad completa entre los sujetos constituyentes (Husserl), o entre el hablante y su interlocutor (Craig);

¹³ E. Husserl, *ibid.*, secc. 52, p. 144.

y tal reciprocidad sólo es posible, a su vez, suponiendo que se cuenta con una objetivación de esas perspectivas en el mundo social común. Pero este conocimiento de las distintas perspectivas sociales, susceptibles de objetivación en los plexos de significado de un lenguaje público, es lo que ni Husserl ni Craig pueden suponer en el punto de partida, pues es lo que se pretende llegar a constituir. En sus reconstrucciones de la intersubjetividad hay, por tanto, una petición de principio que no parece salvable.

III

El problema que plantea la discusión de Craig para una concepción pragmatista del significado puede reformularse en términos kantianos: al igual que, para Kant, la conciencia del objeto de conocimiento era al mismo tiempo, inevitablemente y siempre, conciencia de sí -y este es el punto más elevado de esa Filosofía Crítica: el de la *apercepción transcendental*- , para una concepción *intersubjetivista* del significado la comprensión de lo que una expresión significa tiene que ser inevitablemente, y siempre, *conocimiento* de en qué consiste la aplicación correcta de la expresión -lo que sería explícitamente recuperable en el contexto de lo que se ha llamado *actividades normativas*: situaciones de justificación de lo dicho, de revisión crítica de pretensiones puestas en cuestión, de aclaración de conceptos, de enseñanza de usos lingüísticos, y otras prácticas análogas-. El saber del significado iría así ligado al conocimiento de qué cuenta como justificación suficiente de lo dicho, y esta justificación sería diferente según el tipo de uso del lenguaje -según se tratara, por ejemplo, de un discurso teórico o de un discurso práctico-

El problema puesto de manifiesto por la perspectiva intencionalista del solipsismo metodológico -por mucho que sea acorde a intuiciones preteóricas de los hablantes- reside en los supuestos tácitos a que obliga: la supuesta “intercambiabilidad de perspectivas”, metáfora geométrica, es insostenible cuando no se está tratando de dar expresión a un discurso teórico sobre un mundo objetivo de estados de cosas, sino a un discurso práctico sobre el mundo humano de normas y valores, constituido por relaciones interpersonales. Y, por otra parte, para la reconstrucción intencionalista se precisa, en algún momento, suponer una interpretación de las manifestaciones estables del otro, que dejan de ser vistas como meros movimientos físicos y sonidos arbitrarios, para interpretarse como *símbolos*, como parte de un lenguaje signifiante común: pero esto es, precisamente, algo de lo que todavía no se puede tener noción ni experiencia de ningún tipo.

La defensa de Craig, en este punto, consiste en apelar a algo que considera factual, una cuestión de hecho; en efecto, al comienzo del trabajo que se comenta, declara: “Se ha venido sosteniendo, en general, que determinados estados de las criaturas sentientes son privados, en el sentido técnico de que sólo el sujeto que los experimenta puede conocer su naturaleza (...) no es propósito de este capítulo entrar en [estas cuestiones] en profundidad; de momento, el lector ha de asumir -o, al menos, ha de estar dispuesto a aceptar la hipótesis- que hay *verdaderamente* estados que son privados en el sentido definido”¹⁴. Sin embargo, la aparente naturalidad con que Craig lo presenta no debe dejarnos engañar; pues al inicio de la presente discusión se recogía una observación de C. Wright que permitía poner de manifiesto la complejidad de esta hipótesis de Craig: pues el presupuesto de un acceso privilegiado al conocimiento de los propios EIPs hay que verlo, decía Wright, como un *principio constituyente* del propio estado intencional. Esto sólo parece poder entenderse del siguiente modo. Ese autoconocimiento, o autoconciencia -el punto supremo de la *apercepción transcendental* de Kant, el de la conciencia de sí-, no es una cuestión “de hecho”, sino que tiene valor *normativo* y categorial: introduce la propia categoría que contribuye a constituir, aquí la de *estado intencional*, epistémicamente privado. De tal modo que la autoconciencia sería condición para poder decir, ante un estado biológico o neurológico fáctico, que se trata de un estado de conciencia.

Pero la misma dimensión normativa estaría presente en el tipo de teoría intersubjetivista del significado que aquí se está queriendo esbozar: el presupuesto de la identidad intersubjetiva del significado tendría un valor normativo y categorial, en el sentido de que sería principio constituyente de la propia noción de significado. Sólo así es posible dar el paso desde un sistema de señales público, que permita la interacción con éxito, a lo que representa un lenguaje *compartido*.

Cristina Corredor
Univ. de Valladolid
e-mail: corredor@fyl.uva.es

¹⁴ Craig (1991), *ibid.*, p. 127.

PRESUPOSICIÓN E IMPLICATURA CONVENCIONAL

AMBRÓS DOMINGO BELANDO
Universitat de Barcelona

ABSTRACT: The notions of presupposition and implicature are controversial. Both of them belong to the domain of what is communicated with the utterance of a sentence, besides what the sentence uttered explicitly means, but problems arise when we try to define the phenomena more precisely. Some authors believe that the presupposition and the conventional implicature are two separate phenomena, while others maintain that they are exactly the same thing. In this contribution I will argue that they are two distinct phenomena and that the anaphoric character of the presupposition is the cause of the distinction.

KEYWORDS: presupposition, conventional implicature.

El propósito de esta comunicación es sostener la existencia de una distinción entre las nociones de presuposición y de implicatura convencional. Ambas nociones refieren a proposiciones que son comunicadas de forma no explícita mediante la proferencia de un enunciado. El interés en distinguir entre estas dos nociones radica precisamente en que hay quien sostiene que no existe una tal distinción y que presuposición e implicatura convencional no son sino un mismo fenómeno.¹

En primer lugar (I) se presentará someramente la distinción entre lo dicho explícitamente y lo implicado mediante la proferencia de un enunciado. A continuación (II y III) se caracterizarán las nociones de presuposición y de implicatura convencional para finalmente (IV) establecer la distinción entre ambas nociones, no en base a la cancelabilidad de las presuposiciones frente a la no cancelabilidad de las implicaturas convencionales, como algunos sostienen, sino en base al carácter anafórico de la presuposición.

¹ Tal posición se sostiene, por ejemplo, en Karttunen y Peters, 1979.

I

Cuando hablamos a menudo comunicamos a nuestros interlocutores más información de la que literalmente expresan nuestras palabras. Así, a veces decimos “tengo frío” cuando lo que queremos comunicar además de que tenemos frío es “quiero que cierres la ventana”, o decimos “hoy he vuelto a perder al póker” cuando queremos comunicar, además de que hoy hemos perdido al póker, que hemos perdido al póker alguna otra vez antes de hoy, o proferimos “Enrique es político, pero no es mentiroso” cuando queremos decir que Enrique es político y no es mentiroso, y pretendemos además comunicar que en general los políticos son mentirosos y que hay, por tanto, un cierto contraste entre ser político y no ser mentiroso. Decimos que esta información así comunicada (“quiero que cierres la ventana”, “he perdido al póker alguna otra vez antes de hoy”, “hay un cierto contraste entre ser político y no ser mentiroso”) no forma parte de lo explícitamente dicho con nuestra preferencia del enunciado en cuestión, sino que está meramente implicada por dicha preferencia.

La distinción entre lo dicho y lo implicado la encontramos en los escritos de Paul Grice.² En rigor, Grice distingue tres niveles en lo que concierne al análisis del significado comunicado mediante el acto de proferir un enunciado: lo que las palabras del enunciado significan convencionalmente, lo que decimos con estas palabras cuando usamos el enunciado y lo que implicamos con nuestra preferencia. Para nuestros propósitos es relevante la distinción entre el segundo y el tercero de los niveles, entre lo dicho y lo implicado. De acuerdo con la distinción griceana, lo dicho mediante la preferencia de un enunciado es la proposición que con este enunciado se expresa. Lo implicado con esa misma preferencia son las proposiciones no afirmadas explícitamente —no correlacionables con los ítems léxicos que forman el enunciado aseverado—, pero reconocidas tanto por el hablante como por el oyente como parte del contenido total comunicado mediante la preferencia. Así, en los ejemplos anteriores, las proposiciones “quiero que cierres la ventana”, “hoy he vuelto a perder al póker” y “hay un cierto contraste entre ser político y no ser mentiroso” son proposiciones implicadas por las preferencias de los enunciados “tengo frío”, “hoy he vuelto a perder al póker” y “Enrique es político, pero no es mentiroso”, respectivamente.

Conviene hacer aquí una observación al respecto de la noción de implicación. El tipo de implicación al que nos estamos refiriendo no se trata de una implicación lógica, y no lo es al menos por tres razones. La primera, porque para calcular las proposiciones de este modo implicadas no

² Grice, 1967.

procederemos de acuerdo con las reglas del cálculo de la lógica clásica. En realidad dicho cálculo variará según de qué tipo de fenómeno se trate (implicatura convencional, implicatura conversacional, etc.) y dependerá en algunas ocasiones de aspectos meramente semánticos mientras que en otras será necesario considerar también aspectos pragmáticos. La segunda, porque el comportamiento de la implicación tampoco es el mismo que el de la implicación lógica. En particular, la negación de una proposición implicada de este modo no siempre comportará, como en el caso de la implicación lógica, la negación por *modus tollens* de la proposición a partir de la cual se había generado la implicación. De los tres ejemplos propuestos (“tengo frío”, “hoy he vuelto a perder al póker”, “Enrique es político, pero no es mentiroso”), en el primero la negación de la proposición implicada no afecta al significado del enunciado que la genera ni a sus condiciones de verdad, aunque en el segundo y en el tercero sí —determina cuanto menos que la preferencia de dichos enunciados sea inadecuada. A esto último nos referiremos más adelante cuando consideremos el rasgo distintivo de la cancelabilidad. La tercera, porque a diferencia de las implicaciones lógicas, aquí en algunos casos tanto la preferencia de un enunciado como la de su negación implican la misma proposición. Si retomamos el segundo de nuestros ejemplos, tanto la preferencia de “hoy he vuelto a perder al póker” como la de “hoy no he vuelto a perder al póker” implican “he perdido al póker alguna otra vez antes de hoy”.

Pues bien, como ya se ha apuntado más arriba, presuponer e implicar convencionalmente son dos de las formas según las cuales se genera el contenido implicado por una preferencia.

II

La noción de presuposición responde a la siguiente idea: a veces, cuando hablamos, para interpretar adecuadamente lo que decimos debe suceder que alguna otra proposición que no expresamos sea verdadera. Si consideramos el segundo de los dos ejemplos anteriores, cuando proferimos “hoy he vuelto a perder al póker” la proposición “he perdido al póker alguna otra vez antes de hoy” no se afirma de forma explícita en lo dicho, pero está implicada por nuestra preferencia. En este caso decimos que nuestra preferencia asume dicha proposición, la *presupone*. De no ser así (si fuera falso que hemos perdido alguna otra vez al póker antes de hoy), decir “hoy he vuelto a perder al póker” sería cuanto menos inadecuado, si no sin sentido, y en su lugar deberíamos haber dicho simplemente “hoy he perdido al póker”.

Estas proposiciones presupuestas son evidenciadas por lo general por la presencia de determinados ítems léxicos o por algunas construcciones sintácticas características en el enunciado proferido. Así, encontramos presuposiciones puestas de manifiesto por verbos iterativos (como en nuestro ejemplo con el verbo “volver a”), por verbos factivos (como “lamentar”: proferir “lamento haber venido” presupone que he venido), por construcciones clausales (proferir “fue Fleming quien descubrió la penicilina” presupone que alguien descubrió la penicilina), por el adverbio “también” (proferir “el jefe también vino a la fiesta” presupone que alguien más además del jefe vino a la fiesta, o que el jefe además de a la fiesta fue a algún otro lugar), etc.

A pesar de lo aparentemente intuitivo de este *presuponer* evidenciado en los ejemplos del párrafo anterior, no hay consenso en lo que concierne a la caracterización de la noción de presuposición —incluso, como se advirtió al principio, algunos dudan de la existencia de un tal fenómeno y lo asimilan al de las implicaturas convencionales. Entre quienes defienden la existencia de presuposiciones podemos distinguir básicamente dos líneas de análisis, según se consideren las presuposiciones como un fenómeno semántico o como un fenómeno pragmático. Brevemente, según la primera línea de análisis las presuposiciones son presuposiciones semánticas porque su verdad es necesaria para que una proferencia del enunciado que las genera sea verdadera o falsa. Según el segundo tipo de análisis, las presuposiciones son presuposiciones pragmáticas por cuanto son proposiciones que los participantes en un proceso comunicativo asumen como verdaderas para que sus proferencias sean adecuadas y no dependen sólo, por tanto, de los enunciados finalmente proferidos, sino también y primordialmente de las intenciones comunicativas de los participantes en dicho proceso comunicativo y del contexto en el que tiene lugar la conversación.

No es propósito de esta comunicación dirimir entre cuál de las dos caracterizaciones bosquejadas es la más adecuada. Sin embargo fijémonos en algo que de común hay en ambas, a saber, el compartir la noción intuitiva de *presuponer* consignada más arriba: que deba suceder que alguna otra proposición que no expresamos sea verdadera. Así pues, resulta que cuando la proferencia de un enunciado p (o si se quiere el hablante que profiere p) presupone una proposición q es necesario que q forme ya parte del contexto conversacional en el que se profiere p o se asuma en dicho contexto como verdadera. En la caracterización de presuposición semántica éste es un requisito de su misma definición (q debe ser verdad para que p sea verdadera o falsa); en la noción de presuposición pragmática la verdad de q es asumida por

los participantes en una conversación para poder interpretar adecuadamente la preferencia del enunciado *p*.

La necesidad de que la proposición presupuesta forme parte del contexto conversacional es tal que, si no se halla en dicho contexto, por lo general la añadimos con el fin de interpretar adecuadamente la preferencia que la genera. Si proferimos “hoy he vuelto a perder al póker” en un contexto en el que la proposición “he perdido al póker alguna otra vez antes de hoy” no está presente, procedemos a incrementar el contexto con dicha proposición para poder interpretar de forma adecuada la preferencia. Este hecho de asumir como verdadera en el contexto conversacional una proposición que en realidad no formaba parte de dicho contexto se denomina “proceso de acomodación” —y es difícilmente explicable si consideramos las presuposiciones como un fenómeno meramente semántico.

Debemos, no obstante, tener en cuenta que este proceso de acomodación no se da tan a menudo como pudiera parecer, porque normalmente no forzamos a nuestros interlocutores a un tal esfuerzo acomodaticio para poder interpretar adecuadamente nuestras preferencias. Así, y a pesar de que nada hay de equivocado —aunque sí de extraño— en ello, normalmente no decimos “hoy he vuelto a perder al póker” si no es ya parte del contexto conversacional que hemos perdido al póker algún otro día antes de hoy. Por otra parte, este proceso de acomodación tampoco puede llevarse a cabo si negación de la proposición que deberíamos acomodar está explícitamente negada en el contexto. Así, es inadecuado decir “hoy he vuelto a perder al póker” en un contexto en el que es manifiesto que nunca antes hemos perdido —quizás porque nunca antes habíamos jugado.

A la luz de lo referido hasta aquí, podemos ahora establecer las siguientes condiciones necesarias para caracterizar una presuposición:

un hablante *H* presupone *q* profiriendo un enunciado *p* a un oyente *O*

si:

- (i) *H* dice que *p* (a *O*);
- (ii) *H* no dice que *q* (a *O*);
- (iii) es necesario que *q* forme parte del contexto conversacional —o que se pueda acomodar a dicho contexto— para que la preferencia de *p* exprese una proposición y pueda ser interpretada de forma adecuada por *O*;
- (iv) *H* asume (iii);
- (v) *H* asume que *O* asume (iii) y (iv).

III

“Implicatura” es el neologismo introducido por Grice³ para referirse a aquellas proposiciones que son implicadas —pero no son implicaciones lógicas— por la preferencia de un enunciado. El neologismo tiene precisamente como finalidad evitar el término “implicación”, ligado a la noción de implicación lógica. Grice distingue entre dos tipos de implicatura: implicatura convencional e implicatura conversacional. La diferencia fundamental entre ambos tipos de implicatura es que mientras que las implicaturas convencionales dependen directamente de alguno de los ítems léxicos que conforman el enunciado aseverado, las implicaturas conversacionales no. Dicho de otro modo, en el caso de las implicaturas conversacionales no es posible hallar dos enunciados distintos cuyas preferencias expresen proposiciones que tengan las mismas condiciones de verdad y que no generen la misma implicatura conversacional, contrariamente a lo que sucede en el caso de las implicaturas convencionales —precisamente porque éstas dependen de ítems léxicos concretos. Este rasgo distintivo entre implicaturas convencionales e implicaturas conversacionales se ha convenido en denominar “separabilidad”. Las implicaturas convencionales son separables mientras que las implicaturas conversacionales no lo son.

Retomemos el primer ejemplo expuesto. En lugar de proferir “tengo frío” con la intención de comunicar “quiero que cierres la ventana”, podríamos muy bien haber proferido el enunciado “experimento helor” o cualquier otro que expresara una proposición con las mismas condiciones de verdad que el primero. En este caso la información así implicada (“quiero que cierres la ventana”) se generaría igualmente, de donde se sigue que la implicatura en cuestión es una implicatura conversacional porque no depende de ítems léxicos específicos, sino de lo que se dice con la preferencia, esto es, de la proposición expresada. Por contra, si decimos “es inglés, por lo tanto es valeroso”⁴ con nuestra preferencia implicamos la proposición “ser valeroso es consecuencia de ser inglés”. Dado que resulta factible encontrar un enunciado cuya preferencia exprese una proposición que tenga las mismas condiciones de verdad que la anterior, pero que no genere la implicación en cuestión (“ser valeroso es consecuencia de ser inglés”), por ejemplo, decir “es inglés y es

³ Grice, 1967.

⁴ “He is an Englishman; he is, therefore, brave”. El ejemplo es de Grice (Grice, 1967; 1989, pág. 25 y siguientes). Otro ejemplo de Grice para ejemplificar las implicaturas convencionales es la conjunción adversativa “pero” (Grice, 1967; 1989, pág. 88). Así, una preferencia del enunciado “es pobre, pero honrado” genera como implicatura convencional la proposición “hay un cierto contraste entre ser pobre y ser honrado”.

valeroso”, resulta que la implicación depende de unos ítems léxicos concretos, “por lo tanto”, y se trata, pues, de una implicatura convencional.

Dejemos de lado las implicaturas conversacionales y centrémonos en el caso de las implicaturas convencionales, que es el que aquí nos atañe. De acuerdo con la descripción del fenómeno que hemos dado, podemos ahora establecer las siguientes condiciones necesarias para caracterizar la noción de implicatura convencional:

un hablante *H* implicatura convencionalmente que *q* a un oyente *O* si y sólo si:

- (i) *H* dice que *p* (a *O*);
- (ii) *H* no dice que *q* (a *O*);
- (iii) diciendo que *p* (a *O*) *H* cree que *O* calculará *q* sólo en virtud del significado convencional de algunos ítems léxicos de *p*, independientemente de otra información presente en el contexto conversacional;
- (iv) *H* asume que *O* asume (iii).

IV

Pretender distinguir entre presuposición e implicatura convencional es pretender recoger una distinción en el cálculo de las proposiciones implicadas por las preferencias de enunciados como “hoy he vuelto a perder al póker” y “Enrique es político, pero no es mentiroso”, respectivamente.

Entre quienes sostienen que las presuposiciones y las implicaturas convencionales son efectivamente dos fenómenos distintos, un criterio para efectuar tal distinción se basa en la cancelabilidad⁵. Tradicionalmente se entiende por cancelabilidad el hecho de que una proposición implicada por la preferencia de un enunciado pueda a continuación refutarse (cancelarse) sin alterar por ello las condiciones de verdad del enunciado que la ha generado. Así pues, la proposición “quiero que cierres la ventana” implicada por nuestra preferencia del enunciado “tengo frío” puede ser cancelada si a continuación añadimos “pero no quiero que cierres la ventana”, y ello sin que varíen las condiciones de verdad de nuestra primera preferencia. Éste es un ejemplo de implicatura conversacional, y todas las implicaturas conversacionales son cancelables —porque no dependen de ítems léxicos específicos.

De forma análoga, se argumenta, las presuposiciones son también cancelables mientras que las implicaturas convencionales no lo son. Los casos en los que piensan quienes defienden esta posición son los casos de enunciados compuestos, en donde está en cuestión si las presuposiciones generadas por

⁵ Por ejemplo en Soames, 1989.

separado por los distintos enunciados simples que constituyen el enunciado compuesto se mantienen o no en la preferencia del enunciado mayor. Esto último es conocido como el problema de la proyección. De este modo proferir “me doy cuenta de que mi médico me engaña” presupone que uno es efectivamente engañado por su médico, mientras que proferir ese mismo enunciado en el antecedente de un condicional puede no conllevar tal presuposición: “si me doy cuenta de que mi médico me engaña, no confiaré más en él”.

Sin embargo la cancelabilidad no puede constituir un buen criterio para distinguir entre presuposiciones e implicaturas convencionales, porque en realidad ni las presuposiciones ni las implicaturas convencionales son cancelables en sentido estricto (y si lo fueran lo serían también ambas en la misma medida, como se verá más adelante, en un sentido de cancelabilidad no estricto). Lo que mueve a caracterizar las implicaturas convencionales como no cancelables es su completa dependencia del significado convencional de alguno de los ítems léxicos del enunciado proferido, de manera que pretender cancelar la implicatura convencional comportaría contravenir la presencia en el enunciado proferido del ítem léxico que la genera. Proferir “Enrique es político, pero no es mentiroso” en un contexto en el que sea manifiesto que los políticos nunca mienten determina que el uso de la conjunción adversativa “pero”, que es la que origina la implicatura convencional, sea inapropiado. En su lugar deberíamos haber dicho “Enrique es político y, por lo tanto, nunca miente”. Pero exactamente lo mismo sucede en el caso de las presuposiciones, pues su generación también está originada por la presencia de ítems léxicos determinados o por la de construcciones sintácticas características. Así pues, mantener el enunciado que genera la presuposición cuando querriamos que esta última se cancelara comportaría el mismo tipo de problemas que encontrábamos en el caso de las implicaturas convencionales. Proferir “hoy he vuelto a perder al póker” en un contexto en el que no se encuentre la proposición “he perdido al póker alguna otra vez antes de hoy” ni tampoco sea posible acomodarla, o en un contexto neutro al respecto al que añadiríamos la preferencia del enunciado “nunca antes de hoy había perdido yo al póker”, resultaría tan inapropiado como sucedía en el caso anterior de implicatura convencional.

Si nos enfrentamos al problema de la proyección, la cancelabilidad origina otro tipo de dificultades. Ciertamente es que no todas las proposiciones implicadas por la preferencia de algunos enunciados son implicadas a su vez (y, por tanto, heredadas) por la preferencia de enunciados compuestos que incluyen a los primeros, pero ello tampoco comporta que dichas proposiciones

se cancelen en sentido estricto. La razón es que cuando decimos que alguna de estas proposiciones así implicadas se ha cancelado, lo que sucede también es que la verdad o falsedad de la proposición que la ha generado —y que ahora forma parte de un enunciado compuesto— está, por así decirlo, en suspensión, esto es, no podemos decidir que sea verdadera o que sea falsa. Recordemos que hemos convenido en que la cancelabilidad consiste en el hecho de que una proposición implicada por la proferencia de un enunciado pueda a continuación refutarse (cancelarse) sin alterar por ello las condiciones de verdad del enunciado que la ha generado. Dado que las condiciones de verdad del enunciado en cuestión no son las mismas cuando éste se halla dentro de un enunciado compuesto, resulta, pues, que en realidad no se ha producido una cancelación en sentido estricto. Así, decíamos que proferir “me doy cuenta de que mi médico me engaña” suponía que uno es efectivamente engañado por su médico, mientras que proferir ese mismo enunciado en el antecedente de un condicional no conllevaba tal presuposición: “si me doy cuenta de que mi médico me engaña, no confiaré más en él”. Pero observemos que en este caso el enunciado que genera la posible presuposición está en el antecedente del condicional y, por tanto, su verdad o falsedad está en suspensión. Podemos querer admitir que esta cancelabilidad en sentido no estricto es, pese a todo, cancelabilidad al fin y al cabo; pero entonces conviene observar que lo mismo que sucede con las presuposiciones en enunciados compuestos ocurre también en el caso de las implicaturas convencionales. Una proferencia del enunciado “Enrique es político, pero no es mentiroso” *implicatura* convencionalmente que hay un contraste entre ser político y no ser mentiroso, pero proferir “Juan cree que Enrique es político, pero no es mentiroso, porque [Juan] no sabe que los políticos nunca mienten”, no genera tal implicatura. Así pues, tampoco en el caso del problema de la proyección la cancelación en sentido no estricto nos permite separar entre presuposición e implicatura convencional.

En cualquier caso, la estrategia de reducir la distinción entre los dos fenómenos a su carácter cancelable está, a mi juicio, desencaminada, porque olvida precisamente lo que debe constituir el rasgo determinante de la distinción que perseguimos: el *pre-* de “presuponer” recogido en el punto (iii) de la caracterización de presuposición, que se repite aquí:

- (iii) es necesario que *q* forme parte del contexto conversacional —o que se pueda acomodar a dicho contexto— para que la proferencia de *p* exprese una proposición y pueda ser interpretada de forma adecuada por *O*.

Esta necesidad de la presencia de la proposición presupuesta en el contexto —o la posibilidad de su acomodación— para poder interpretar de

forma adecuada la preferencia del enunciado que decimos genera la presuposición es lo que separa presuposición de implicatura convencional y es, a su vez, lo que podríamos proponer a la manera de un test: cuando la proposición implicada lo sea a partir de construcciones o de ítems léxicos determinados, si es necesario asumir la proposición implicada para interpretar adecuadamente la preferencia, entonces se trata de una presuposición. Si por contra la proposición implicada se añade al significado explícito de la preferencia en tanto que información adicional, pero aunque no forme parte del contexto conversacional la preferencia no resulta en absoluto inadecuada, entonces se trata de una implicatura convencional. Dado que este rasgo establece el cálculo de las presuposiciones vinculando directamente algún ítem léxico o construcción sintáctica con alguna proposición presente ya (o acomodable) en el contexto conversacional, podemos denominarlo “anaforicidad”, para concluir afirmando que las presuposiciones tienen este carácter anafórico mientras que las implicaturas convencionales no.

Retomemos por última vez dos de nuestros ejemplos. Cuando decimos “hoy he vuelto a perder al póker” nuestra preferencia expresa una proposición siempre y cuando, o bien la proposición “he perdido al póker alguna otra vez antes de hoy” esté explícitamente en el contexto, o bien podamos acomodarla para interpretar adecuadamente la preferencia, en particular la construcción verbal iterativa “volver a perder”. Se trata, pues, de una presuposición. En cambio proferir “Enrique es político, pero no es mentiroso” no requiere que la proposición “hay un cierto contraste entre ser político y no ser mentiroso” forme parte del contexto ni siquiera que deba previamente acomodarse a él para que nuestra preferencia resulte adecuada. Se trata, pues, de una implicatura convencional.

Hasta donde conozco, el primero en vincular la noción de anaforicidad al fenómeno de las presuposiciones fue Saul Kripke.⁶ No obstante, y a pesar de lo acertado de tal vínculo, los ejemplos que sugiere Kripke de presuposición con anáfora resultan ser claros casos de implicaturas conversacionales. Posteriormente Rob van der Sandt⁷ ha propuesto un análisis de las presuposiciones como anáforas dentro del marco de la Teoría del Discurso Representacional.

BIBLIOGRAFÍA

Grice, Paul: “Logic and Conversation”, 1967, en el capítulo 2 de Grice, 1989.

Grice, Paul: *Studies in the Way of Words*, Cambridge, MA.: Harvard University Press, 1989.

⁶ Kripke, 1991.

⁷ Sandt, 1992.

- Karttunen, Lauri y Peters, Stanley: "Conventional implicature", en *Syntax and Semantics*, volumen 11, 1979.
- Kripke, Saul: "Presupposition and Anaphora: remarks on the formulation of the projection problem", transcripción de una conferencia dada en Princeton, fotocopia, 1991.
- Sandt, Rob v.d.: "Presupposition Projection as Anaphora Resolution", en *Journal of Semantics*, 9, 1992.
- Soames, Scott: "Presupposition", en el *Handbook of Philosophical Logic*, editado por D. Gabbay y F. Guentner, volumen IV, D. Reidel Publishing Company, 1989.
- Stalnaker, Robert: "Pragmatic Presuppositions", en *Semantics and Philosophy*, editado por Milton K. Munitz y Peter K. Unger, New York University Press, 1974.

Ambròs Domingo Belando
Seminari de Filosofia - IES Torredembarra
e-mail: adb@tinet.fut.es





ALGUNOS ASPECTOS FILOSÓFICOS Y COMPUTACIONALES DE LA SINONIMIA*

SANTIAGO FERNÁNDEZ LANZA
Universidad de Santiago de Compostela
ALEJANDRO SOBRINO CERDEIRIÑA
Universidad de Santiago de Compostela

ABSTRACT: It is frequently said that two expressions are synonyms if they have the same meaning. But meaning is a polysemic word. This work explores different definitions of 'meaning', under the hypothesis that, if we want to consider synonymy as a crisp relation defined in terms of meaning, we must accept a structural conception of it. In this vein, the meaning of an expression should be characterised by the admitted transformations with respect to a task. Considered as a crisp relation, synonymy presupposes that the meaning of an expression is a truth value and the task is to propagate truth in inferential chains. Logic programming is a useful tool for studying transformations between expressions preserving truth. A Prolog program for the passive case illustrates reasonably our hypothesis.

KEYWORDS: Synonymy, meaning, natural language processing.

1. Preliminares

Con frecuencia, se dice que la sinonimia es una característica propia de las expresiones de los lenguajes naturales y, si estas expresiones son palabras, una definición informal de sinonimia suele indicar que dos palabras son sinónimas si significan lo mismo o aproximadamente lo mismo. Otras veces, la sinonimia se predica de oraciones: dado que las oraciones se componen de palabras, la sustitución de algunas palabras por sus sinónimas conforman oraciones también sinónimas; esto es, con igual o parecido significado al que tiene la oración original.

Dado que la sinonimia es una característica que atañe, en último término, al significado de las palabras, parece oportuno debatir tan polémica cuestión. Si la sinonimia se define por la igualdad o similitud de significado, preguntar qué es la sinonimia presupone, en cierto modo, contestar a qué es el significado. En este trabajo no se pretende responder a esta cuestión, sino sólo

* Financiación a cargo del proyecto PGIDT99PXI10502B de la Xunta de Galicia.

examinar distintas aproximaciones a ella y justificar la elección de alguna de las alternativas para un tratamiento automático de la sinonimia.

El trabajo se estructura de la siguiente manera. En el apartado 2 se contrapone la semántica de los lenguajes formales a la semántica de los lenguajes naturales y se indica cómo la lógica formal ha servido para regimentar los significados del lenguaje común. En la sección 3 se analizan algunos problemas filosóficos ilustrados con la sinonimia y se pone de manifiesto las contradicciones que surgen por querer tratar al lenguaje natural, de naturaleza imprecisa, de modo preciso. En el apartado 4 se propone analizar la sinonimia entre palabras en términos de transformación, examinando las alteraciones de palabras que preservan valor de verdad oracional. En la sección siguiente se propone un programa Prolog para estudiar la transformación de pasiva, que vale para ilustrar la sinonimia transformacional. En la última sección, se apunta a otra concepción posible de la sinonimia, la sinonimia empírica y algunos para tratarla.

2. Semántica y significado

Al ámbito de la teoría del lenguaje que estudia el significado se le conoce como semántica. En el apartado que sigue aludiremos, al comentar la sinonimia, a distintos tipos de semántica, pues lejos de ser un ámbito preciso, es plural.

Es posible distinguir al menos dos acepciones diferentes de la palabra ‘semántica’, dependiendo, p. ej., si se usa en el contexto de un lenguaje preciso o impreciso. Habitualmente, el primer caso se ejemplifica en la semántica de un lenguaje formal y, el segundo, en la semántica de un lenguaje natural. Para abreviar, al primero le llamaremos ‘semántica_{LF}’ y al segundo ‘semántica_{LN}’.

El objetivo de la semántica_{LF} es establecer una correspondencia biunívoca entre expresiones lingüísticas, de manera que no quepa ambigüedad o vaguedad alguna sobre qué significa cada símbolo del lenguaje formal. Cada uno de sus símbolos referirá a un símbolo o cadena de símbolos del lenguaje objeto, del que se presupone que a su vez denotará, de una manera transparente y nítida, a alguna realidad que mostrará el lenguaje. Así pues, en una semántica formal, el significado se entiende como la materialización de una expresión lingüística en el sistema de lenguaje considerado. Dado que lo que se establece es una relación entre entidades lingüísticas, esta semántica se ocupa, preferentemente, de expresar verdades analíticas.

La semántica_{LN} se ocupa principalmente, en cambio, de verdades contingentes, de oraciones cuyo significado no resulta de la forma que exhiben las palabras en la oración, sino de su referencia a entidades extralingüísticas, como estados de cosas o cogniciones. Bajo esta aproximación, el significado aparece frecuentemente como algo oscuro y no dependiente sólo de capacidades razonadoras y argumentativas, sino de disposiciones físicas, como la memoria y la observación. Para desvelar la oscuridad del lenguaje natural se

ha recurrido a los lenguajes formales, que tratan de traducir a un lenguaje preciso lo que de manera confusa parece expresar la lengua común. Al hacerlo así, los lenguajes formales no sólo aspiran a ser una notable empresa intelectual, sino también una ciencia útil. Pero de ese mestizaje le sobrevienen muchos problemas.

A menudo, los lenguajes naturales -entre los que se suele incluir al lenguaje científico- han sido objeto de estudio por parte de los lenguajes formales. Una precondition para aplicar la semántica_{LF} a las oraciones del lenguaje natural es que éstas se hayan sometido previamente a un proceso de regimentación. La regimentación consiste en crear las condiciones de conocimiento o de contexto adecuadas para que una oración usualmente vaga se transforme en una oración precisa. Se trata de traducir o parafrasear una oración de un lenguaje oscuro y ambiguo, a otro claro y determinado, que entonces se mostrará semánticamente transparente, para desvelar su estructura formal. Proporcionar una traducción o paráfrasis conlleva elucidar -en un sentido carnapiano- la oración que se traduce en términos de la oración traducida; esto es, mostrar que el *explicans* es sinónimo del *explicandum* y, en tanto que se deduce de él, es más claro e informativo. La semántica de los lenguajes formales traducen o parafrasean la semántica de los lenguajes naturales, elucidándola. No obstante la traducción de un lenguaje natural a un lenguaje formal presenta dificultades (asociadas a las características de la semántica de un lenguaje natural).

Los lenguajes naturales no traducen ni elucidan, sino que pretenden describir aquello a lo qué refieren las palabras o las oraciones; indicar en vez de qué está la cadena de símbolos empleados para verbalizar algo. Esta tarea, sin embargo, está lejos de ser clara. Usualmente se ha considerado que las palabras o las oraciones pueden significar:

- aquello que se aserta con la oración; esto es, una sucesión de palabras.
- el objeto o situación que denota; la realidad extralingüística a la que refiere el lenguaje.
- la entidad mental asociada con la expresión; simplifícadamente, su 'contenido cognitivo'.

Analicemos la primera alternativa. Parece una propuesta sensata sugerir que aquello que expresa una oración es lo que la oración dice. Por tanto, que es correcto asertar la oración 'Todo profesor estima a sus alumnos' si y sólo si todo profesor estima a sus alumnos, dado que ambas son sinónimas de forma precisa. Pero la tarea desentremolladora facilitada por este criterio parece ofrecer escasa informatividad, por ser tautológica y presuponer que la oración equivalente por el lado derecho es referencialmente translúcida, característica de difícil defensa en el lenguaje natural. Además, con este criterio se le supone al lenguaje una autonomía total en el papel de fijar el significado. Estas

dificultades conducen a una alternativa al significado como expresar: el significado revelado por la traducción.

Se puede sostener razonablemente que comprender el significado de una oración presupone saber cuál es su significado. Asumamos que saber el significado de 'U es P' es describir los $x \in U$ que caen bajo la extensión de P. Pues bien, para poner en claro el significado extensional de la oración 'Todo profesor estima a sus alumnos', que mejor que acudir a un lenguaje formal, como la lógica de predicados de primer orden, para desvelar su significado. Lo conseguiríamos al reescribirla con su ayuda así: 'para todo x e y , si x es profesor e y alumno, x estima a y '. Un problema asociado a esta interpretación es que estaríamos obligados a atribuir el mismo significado a dos oraciones tan dispares como 'todos los gnomos son fantásticos' y 'todas las brujas son fantásticas', al provocar la vacuidad referencial del antecedente la verdad del condicional.

Cabría pensar que las lógicas intensionales son capaces de mejorar una propuesta sobre qué es el significado. Caso de ser capaces de aprehender algún contenido de conciencia independientemente de cualquier articulación lingüística, habría dos tipos de entidades, unas lingüísticas y otras mentales, de manera que unas significarían a las otras. Pero parece difícil describir la entidad mental que está detrás de 'todo profesor estima a sus alumnos'. Quizás en este, como en el caso anterior, podemos sentirnos tentados a reclamar la presencia de un lenguaje formal, en este caso intensional, que nos ayude en tal tarea. Diríamos entonces que, el correlato mental de la oración 'Todo profesor estima a sus alumnos' sería el pensamiento de que para todo x e y si x es profesor e y es alumno, x estima a y '. Pero los pensamientos no tienen entidad lógica independientemente de su verbalización, con lo cual estaríamos en el caso descrito en el apartado anterior. Podríamos intentar recurrir, en vez de a los pensamientos, a sus correlatos neuronales. En ese caso, el significado de 'todo profesor estima a sus alumnos' sería el proceso físico que vincularía, a través de la excitación o estimulación de determinada área cerebral, a todos los profesores y alumnos bajo la propiedad estimatoria de los primeros a los segundos. A diferencia de las entidades mentales, los estados cerebrales gozan, en cierto modo, de autonomía respecto a las oraciones con las que se pueden expresar. Pero si se hace sin relacionarlos con ellas, no nos proporcionan los significados de las proposiciones de pensamiento, sino el significado de las oraciones sustitutas.

Esto indica que la especificación del significado refiriéndolo a expresiones lingüísticas, hechos o actos mentales plantea problemas a los que aquí sólo se ha aludido brevemente. Estos problemas contagian a la sinonimia cuando se estudia en términos de significado. A continuación, aludiremos brevemente a este hecho.

3. Significado y sinonimia

Como ya se indicó, es posible contemplar a la sinonimia desde una perspectiva plural:

- como una relación precisa entre palabras u oraciones (en este caso, se dice que dos palabras u oraciones son sinónimas si significan lo mismo)
- como una relación aproximada entre palabras u oraciones (en este caso, se dice que dos palabras u oraciones son sinónimas si significan lo mismo o aproximadamente lo mismo). Obsérvese como esta última descripción incluye a la primera como un caso particular.

Analicemos algunas características de la sinonimia como relación precisa. Así considerada, la sinonimia se ha estudiado en el ámbito de la semántica formal, usualmente en el marco de una lógica de predicados de primer orden. Una concepción formal de la semántica hace depender el significado de la estructura lógica de la oración, puesta de manifiesto por un análisis lógico que requiere de enunciados asertivos de significado preciso, que puedan ser catalogados como verdaderos o falsos. Por tanto, para estudiar relaciones lógicas interesantes que incluyan a la sinonimia como protagonista principal, deberemos ser capaces de aislar oraciones con significado preciso que sean sinónimas de manera inequívoca; esto es, que tengan el mismo significado. Sin embargo, este tipo de oraciones, más que para examinar a la sinonimia propiamente dicha, se han empleado para estudiar distintos problemas filosóficos porque, dada la circularidad y problematicidad que ofrece un análisis de la sinonimia en términos de significado, los filósofos analíticos han preferido, en vez de discutir directamente acerca de ella, ejemplificar con ella temas filosóficos relevantes. La analiticidad, la intersustituibilidad salva veritate, o la referencia vacua son tres aspectos frecuentemente estudiados en términos de sinonimia.

En el debate filosófico, ha sido usual tratar a la sinonimia como ilustradora de la analiticidad. Así, suele decirse que un enunciado es analítico si

- es una verdad lógica, o
- si es transformable en una verdad lógica sustituyendo algunas de sus expresiones por sus sinónimas.

Ejemplo de verdad lógica es ‘Ningún hombre no saciado está saciado’. Ejemplo de un enunciado que aún no expresando una verdad lógica es analítico, es ‘Ningún hombre hambriento está saciado’, por ser ‘hambriento’ sinónimo de ‘no saciado’. Pero las dificultades para definir lo que es la sinonimia repercuten en la claridad de la analiticidad de la segunda oración, ya que ésta no es analítica por su forma, sino porque se presupone que, al traducirla a un lenguaje lógico, resultarían expresiones iguales. Por tanto, la última oración es analítica porque ‘hambriento’ y ‘no saciado’ son sinónimos; pero a su vez estas dos palabras son sinónimas porque las dos oraciones completas son equivalentes y, como

tales, ambas analíticas, al suponer que ‘habriento’ es la negación perfecta de ‘saciado’. Pero esta argumentación es falaz por circular.

Una forma de superar esta dificultad sería apelar al recurso de la definición, haciendo ‘no saciado’_{def.} ‘hambriento’. La definición debe basarse en la competencia del lexicógrafo, pero su trabajo proporciona datos lingüísticos, no criterios a los que se aspira con la traducción del lenguaje natural al lenguaje formal. La sinonimia como equivalencia definicional es incluso difícil de mantener en un lenguaje formal. P. ej., cabe pensar en un sistema lógico donde $a \rightarrow b$ y $\neg a \vee b$ sean expresiones equivalentes, pero no tengan el mismo significado (no sean, por tanto, sinónimas), porque haya una regla deductiva para $\neg a \vee b$ y no para $a \rightarrow b$. Tal ocurre, p. ej., en el cálculo de resolución de Robinson.

También ha formado parte del debate filosófico la convicción de que la sinonimia entre dos expresiones consiste en la intercambiabilidad de los sinónimos *salva veritate*. A esta tesis se le han puesto objeciones artificiosas, como que siendo ‘soltero’ sinónimo de ‘hombre no casado’, “Soltero’ tiene menos de diez letras’ no es intersustituible por “‘Hombre no casado’ tiene menos de diez letras’, ya que una es verdadera y la otra falsa. Esta objeción se desvanece al notar que en la intersustituibilidad, la sinonimia debe considerarse sobre toda una palabra u oración, no sobre una parte, como es el caso del ejemplo. Una objeción más natural es la basada en la siguiente argumentación. ¿Cuál es el criterio según el cual la oración ‘Ningún hombre habriento está saciado’ se puede sustituir, sin pérdida del valor de verdad, por la oración ‘Ningún hombre no saciado está saciado’?. Que “‘Ningún hombre habriento está saciado’ si y sólo si ‘Ningún hombre no saciado está saciado’” es una oración compuesta y analítica. Pero, ¿por qué es analítica?. Porque ‘habriento’ es sinónimo de ‘no saciado’. Y por qué ‘habriento’ es sinónimo de ‘no saciado’. Porque su empleo en distintas oraciones en las que todo permanezca igual excepto la palabra intercambiada, no altera su valor de verdad. Se puede comprobar, una vez más, la circularidad del razonamiento, que recoge una parte del bucle ya atribuido a la definición de sinonimia en términos de analiticidad.

Cabría entonces pensar en apostar por la extensionalidad, de modo que se dijera que dos palabras (predicados) son sinónimas si tienen la misma extensión. Pero la extensionalidad es condición necesaria, aunque no suficiente para la sinonimia. Ya hemos dicho que, en el caso de los predicados vacuos, este criterio hace sinónimos a ‘gnomos’ y ‘brujas’, porque ambos carecen de extensión, aunque evidentemente, ambas palabras tienen significados diferentes.

Algunos filósofos han tratado de salvar estas objeciones proponiendo nociones alternativas, como las de ‘significado analítico’ (C. I. Lewis) o ‘extensiones primarias y secundarias’ (N. Goodman). No abundaremos en ello. De modo simplificado, se puede decir que estas propuestas no arreglan, sino

sólo parchean, las dificultades aludidas de ejemplificar problemas filosóficos con palabras u oraciones a las que se le supone una sinonimia perfecta.

En el discurso ordinario la sinonimia perfecta no es fácil de encontrar y el corsé de la regimentación que le pretende aplicar un lenguaje formal como la lógica de primer orden se ve desbordado con relativa frecuencia. ‘Saciado’ o ‘hambriento’ son palabras del lenguaje natural de difícil precisión. En estos casos, cuando se intenta aplicar la semántica de un lenguaje formal para sistematizar y pretendidamente hacer transparente la semántica de un lenguaje natural, las dificultades sobrevienen.

¿Cómo acercarse, pues, al estudio de la sinonimia? Si lo que deseamos es estudiar la sinonimia en el ámbito del lenguaje natural, la aproximación ha de ser empírica. Se trata de investigar, respecto a un lenguaje dado, oraciones del tipo ‘Decir A es sólo decir B’, ‘A, en otras palabras, B’ o ‘Cuando digo A sólo quiero decir B’. El uso de estas oraciones suele ser dependiente de las personas y situaciones en las que son utilizadas, cuestión en la que ya reparó J. S. Mill o, contemporáneamente, B. Mates, M.G. White o A. Naess. Según que las personas o situaciones sean iguales o diferentes habrá distintos tipos de sinonimia: intrapersonal, interpersonal, intrasituacional o intersituacional. El hallar el sinónimo o sinónimos de una palabra dada dependerá, por tanto, de una investigación aplicada y requerirá, p. ej., cuestionarios estandarizados que permitan confirmar hipótesis o análisis de textos que extraigan regularidades. Esta investigación revelaría que la sinonimia es una cuestión de grado, ya que unas personas atribuirían, respecto a una situación particular o a una tarea pretendida, una lista de sinónimos para una palabra dada, que podría diferir de otra confeccionada en otra situación o con otro propósito. Pero esta iniciativa, si bien atiende al uso real del lenguaje ordinario, no da un criterio para definir cuándo dos palabras son sinónimas. Si acaso, confirma o contraviene la propuesta del lexicógrafo que se recoge en un diccionario de sinónimos. Y si bien esto es importante, también es fundamental disponer de un método para definir o caracterizar la sinonimia.

Ya se ha dicho que una definición de sinonimia depende de una caracterización del significado. Pero esto, como se vio, presenta problemas si se considera al significado en términos de representaciones mentales, referentes objetuales o expresiones lingüísticas que coincidan o no exactamente con aquello que se quiera significar. ¿Es posible examinar el significado de una manera distinta?. A responder a esta pregunta dedicaremos la siguiente sección.

4. Sinonimia, traducción y transformación

Hasta ahora, la definición de la sinonimia en el marco de un lenguaje formal se hacía depender de un criterio traductor del lenguaje natural a un lenguaje formalizado, que desvelase o pusiese en claro la vaguedad de significados. Como ya se ha dicho, esto ocasiona numerosos bucles de

razonamiento. Pero tales dificultades no deben llevarnos a prescindir del lenguaje formal, sino a hacerlo quizás en otros términos, de manera que la definición de sinonimia sea dependiente, no de una propuesta de traducción, sino de un criterio de transformación. Al fin, la tarea primordial de la lógica no es tanto determinar qué es la verdad, cuanto propagar correctamente la verdad en las inferencias. Desde este punto de vista, el significado de una palabra puede identificarse con el papel que ejerce esa palabra respecto a un propósito determinado, como p. ej., la transmisión de la verdad o la propagación de la gramaticalidad.

De lo que se trata es de explicitar la tarea o tareas para las que se propone el lenguaje, poniendo de relieve el papel de las expresiones individuales en esa empresa. Una semántica formal precisa no debe pretender expresar oraciones del lenguaje natural, señalar los hechos del mundo o los estados cognitivos que le corresponden, sino representar los roles de las palabras y las oraciones dentro de un propósito para el que se diseña el sistema de lenguaje que se considere (p. ej., el propósito inferencial). En este caso, el significado de una palabra u oración queda determinado por aquello que está permitido inferir de ella.

El desarrollo de lenguajes de programación que llevan los roles inferenciales implícitos en su semántica facilita caracterizar al significado de esta manera. Tal es el caso del lenguaje de programación lógica Prolog.

La semántica de la programación lógica trata de reconstruir la semántica del lenguaje natural desde la perspectiva de las deducciones lógicas (automáticas) que posibilita. Por su proximidad al lenguaje natural -con cierta estructura- y al lenguaje formal de la lógica de predicados de primer orden, permite que

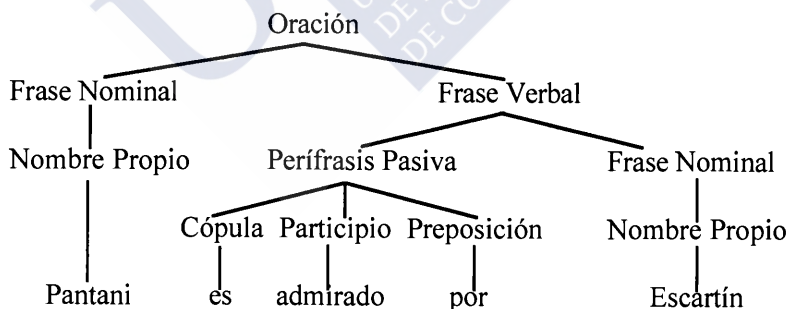
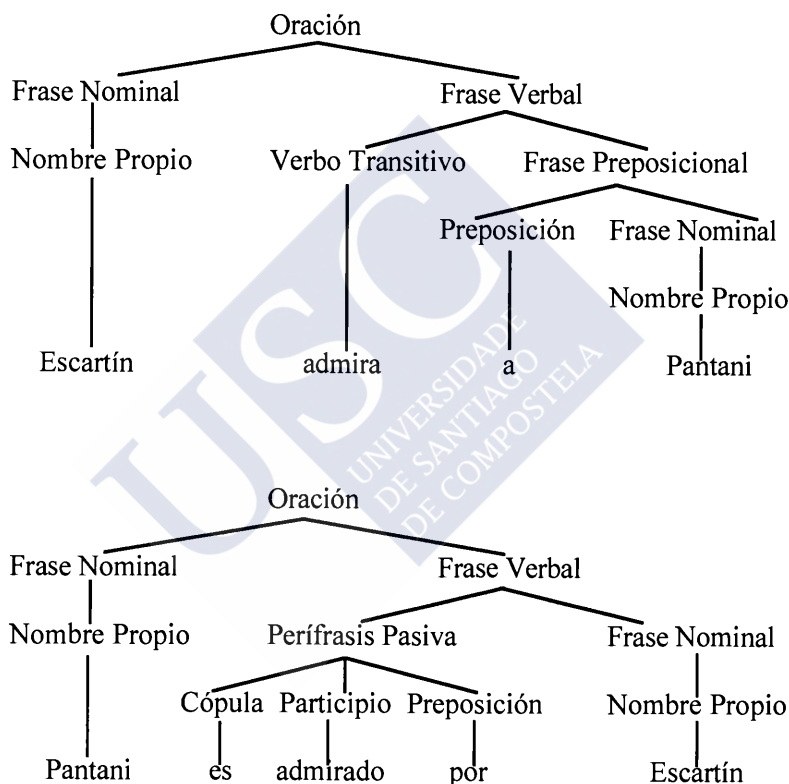
- el papel inferencial del lenguaje natural (explicandum) se aproxime razonablemente al papel inferencial del lenguaje de programación (explicans).
- que el lenguaje de programación sea, respecto a la tarea inferencial, más explícito que el lenguaje natural y, por tanto, un sustituto apropiado.
- que al trabajar con bases de datos, refiera a ámbitos de conocimiento explícitos y sobre cuya verdad tiene que comprometerse el programador, evitando algunas objeciones, como la de la referencia vacua.

De esta manera, en la semántica de un programa lógico los términos no relacionan palabras con cosas, ni vinculan palabras a pensamientos, sino que sirven para desplegar un lenguaje con cierta pre-estructura en otro lenguaje con una estructura más acabada. El papel que tengan en esta tarea los términos vale para determinar su sinonimia. El caso de la transformación de pasiva y su tratamiento en Prolog, ofrece un señalado ejemplo de lo que decimos.

5. Programa prolog para la sinonimia como transformación de pasiva.

5.1. La gramática

Un caso típico de sinonimia entre dos oraciones es la transformación de pasiva. Resulta evidente que la oración “Escartín admira a Pantani” tiene el mismo significado que “Pantani es admirado por Escartín”, aunque se trata de dos oraciones distintas con una estructura gramatical diferente, como se puede comprobar con sus respectivos árboles gramaticales:



Para detectar de forma automática este tipo de sinonimia se puede recurrir a realizar un análisis semántico con técnicas de Prolog. El resultado debería ser un programa Prolog capaz de fijar el significado de una oración en función del significado de sus componentes y del papel que ejercen estos en su estructura sintáctica, mostrando que aunque “Escartín admira a Pantani” y “Pantani es admirado por Escartín” tienen distinta estructura superficial, responden a la misma estructura profunda, al ser transformable, bajo preservación de verdad, una en la otra. El análisis semántico revelaría que el significado de ambas oraciones es *admira_a(Escartín, Pantani)*, aunque para significar esto respondan al uso de distintas reglas.

Fieles al paradigma de la programación lógica, el análisis semántico presupone el análisis sintáctico o gramatical, que lo proporciona, para oraciones como las señaladas, el siguiente programa:

```

oracion(Lista1, Resto):-
    frase_nominal(Lista1, Lista2),
    frase_verbal(Lista2, Resto).
frase_nominal(Lista1, Resto):-
    nombre_propio(Lista1, Resto).
frase_verbal(Lista1, Resto):-
    verbo_transitivo(Lista1, Lista2),
    frase_preposicional(Lista2, Resto).
frase_verbal(Lista1, Resto):-
    perifrasis_pasiva(Lista1, Lista2),
    frase_nominal(Lista2, Resto).
frase_preposicional(Lista1, Resto):-
    preposicion(Lista1, Lista2),
    frase_nominal(Lista2, Resto).
perifrasis_pasiva(Lista1, Resto):-
    copula(Lista1, Lista2),
    participio(Lista2, [por|Resto]).
nombre_propio([escartin|Resto], Resto).
nombre_propio([pantani|Resto], Resto).
verbo_transitivo([admira|Resto], Resto).
preposicion([a|Resto], Resto).
copula([es|Resto], Resto).
participio([admirado|Resto], Resto).

```

A la pregunta de qué oraciones genera:

```
?- oracion(Oracion, []).
```

el compilador responde:

```

Oracion = [escartin,admira,a,escartin]
Oracion = [escartin,admira,a,pantani]
Oracion = [escartin,es,admirado,por,escartin]
...

```

5.2. El análisis del significado con Prolog

Una vez que tenemos la gramática que genera las oraciones que deseamos, debemos incorporar en ella los criterios por los que se debe guiar el análisis semántico.

(a) En primer lugar, el significado de un nombre propio es el individuo denotado por dicho nombre. Esta información debe ser proporcionada al compilador en la regla donde se da cuenta de los nombres propios. Se proporcionará incorporando la palabra empleada para definir al individuo que denota el nombre propio, como primer argumento del predicado `nombre_propio`. Así tendremos:

```

nombre_propio(escartin,[escartin|Resto], Resto).
nombre_propio(pantani,[pantani|Resto], Resto).

```

(b) El significado de un verbo transitivo será una relación, cuyo nombre es la palabra empleada para el verbo, que posee dos argumentos: un agente y un

paciente. Como la información de quién es el agente y quién el paciente viene dada por el significado de otros componentes de la oración, ambos deben ser incorporados también como argumentos del predicado verbo_transitivo (a este proceso se le suele llamar herencia). Por tanto, el agente, el paciente y el significado del verbo transitivo serán los tres primeros argumentos del predicado verbo_transitivo:

```
verbo_transitivo(Agente, Paciente, admira_a(Agente, Paciente),
[admira|Resto], Resto).
```

(c) El significado del participio es el mismo que el de un verbo transitivo:

```
participio(Agente, Paciente, admira_a(Agente, Paciente), [admirado|Resto],
Resto).
```

(d) El significado de la perífrasis de pasiva se hereda del significado del participio, que es uno de sus componentes:

```
perifrasis_pasiva(Agente, Paciente, Sgdo, Lista1, Resto) :-
    copula(Lista1, Lista2),
    participio(Agente, Paciente, Sgdo, Lista2, [por|Resto]).
```

(e) El significado de una frase nominal es el mismo que el de su único componente; nombre propio:

```
frase_nominal(X, Lista1, Resto) :-
    nombre_propio(X, Lista1, Resto).
```

Y el de una frase preposicional es heredado del significado de la frase nominal que es uno de sus componentes:

```
frase_preposicional(X, Lista1, Resto) :-
    preposicion(Lista1, Lista2),
    frase_nominal(X, Lista2, Resto).
```

(f) La mayor dificultad está en la frase verbal.

Si está en activa, el significado de la frase nominal que va después del verbo transitivo es el paciente. En este caso, como es necesaria la información sobre quién es el agente, que se obtendrá del significado de la frase nominal que encabeza la oración, se incluirá dicha información como primer argumento, mientras que el segundo será el significado de la frase verbal, que será la composición del significado del verbo transitivo con el significado de la frase preposicional que lo sigue:

```
frase_verbal(Agente, X, Lista1, Resto) :-
    verbo_transitivo(Agente, Paciente, X, Lista1, Lista2),
    frase_preposicional(Paciente, Lista2, Resto).
```

Si la frase verbal está en pasiva el significado de la frase nominal que va después del verbo transitivo es el agente. En este caso, como es necesaria la información sobre quién es el paciente, que se obtendrá del significado de la frase nominal que encabeza la oración, se incluirá dicha información como primer argumento, mientras que el segundo será el significado de la frase verbal, que será la composición del significado de la perífrasis pasiva con el significado de la frase nominal que lo sigue:

```
frase_verbal(Paciente, X, Lista1, Resto) :-
    perifrasis_pasiva(Agente, Paciente, X, Lista1, Lista2),
    frase_nominal(Agente, Lista2, Resto).
```

Finalmente, el significado de la oración es el significado de la frase verbal compuesto con el significado de la frase nominal que encabeza la oración, que proporciona la información de quién es el agente en el caso de las oraciones en activa y quién es el paciente en el caso de las oraciones en pasiva:

```
oracion(O,Lista1, Resto):-
    frase_nominal(X,Lista1, Lista2),
    frase_verbal(X,O,Lista2, Resto).
```

Los demás componentes de la oración no influyen en el significado tal y como aquí es concebido para este ejemplo concreto; sin embargo, son necesarios para la generación de las oraciones gramaticales. Por tanto, el programa completo sería el siguiente:

```
oracion(O,Lista1, Resto):-
    frase_nominal(X,Lista1, Lista2),
    frase_verbal(X,O,Lista2, Resto).
frase_nominal(X,Lista1, Resto):-
    nombre_propio(X,Lista1, Resto).
frase_verbal(Agente,X,Lista1,Resto):-
    verbo_transitivo(Agente,Paciente,X,Lista1, Lista2),
    frase_preposicional(Paciente,Lista2, Resto).
frase_verbal(Paciente,X,Lista1,Resto):-
    perifrasis_pasiva(Agente,Paciente,X,Lista1,Lista2),
    frase_nominal(Agente,Lista2,Resto).
frase_preposicional(X,Lista1, Resto):-
    preposicion(Lista1, Lista2),
    frase_nominal(X,Lista2, Resto).
perifrasis_pasiva(Agente,Paciente,Sgdo,Lista1, Resto):-
    copula(Lista1, Lista2),
    participio(Agente,Paciente,Sgdo,Lista2, [por|Resto]).
nombre_propio(escartin,[escartin|Resto], Resto).
nombre_propio(pantani,[pantani|Resto], Resto).
verbo_transitivo(Agente,Paciente,admira_a(Agente,Paciente),
[admira|Resto], Resto).
preposicion([a|Resto], Resto).
copula([es|Resto], Resto).
participio(Agente,Paciente,admira_a(Agente,Paciente),
[admirado|Resto], Resto).
```

Un predicado que sea capaz de detectar si dos oraciones *A* y *B* son sinónimas podría ser entonces el que tiene en cuenta que *A* y *B* son gramaticales y su significado es el mismo. Tal predicado tendrá dos argumentos que serán la oración *A* y la oración *B*. Veámoslo:

```
sinonima(A,B):-
    oracion(Significado,A,[]),
    oracion(Significado,B,[]).
```

5.3 Ejemplos de sinonimia entre oraciones

Ahora podremos preguntar al compilador cual es el significado de cada una de las oraciones generadas por la gramática:

```
oracion(Significado,Oracion,[]).
```

A lo que el compilador responde (no incluimos la lista exhaustiva, sólo los ejemplos más ilustrativos)

```
...
Significado = admira_a(escartin, pantani)
Oracion = [escartin, admira, a, pantani]
...
Significado = admira_a(escartin, pantani)
Oracion = [pantani, es, admirado, por, escartin]
...
```

Ahora podemos preguntar si una oración dada es sinónima de otra:

```
?.-sinonima([escartin,admira,a,pantani],[pantani,es,admirado,
por,escartin]).
```

y el compilador responde: yes

También podemos preguntar cuáles de las oraciones generadas por la gramática son sinónimas. Es evidente que toda oración es sinónima de sí misma, hecho del que da cuenta el programa. Si se quieren evitar los casos de sinonimia que incluyan oraciones idénticas se debe formular la pregunta de la siguiente forma:

```
?.- sinonima(Oracion1,Oracion2), Oracion1 \== Oracion2.
```

A lo que el compilador responde (sólo recogemos los ejemplos significativos):

```
...
Oracion1 = [escartin,admira,a,pantani]
Oracion2 = [pantani,es,admirado,por,escartin]
...
Oracion1 = [pantani,es,admirado,por,escartin]
Oracion2 = [escartin,admira,a,pantani]
...
```

6. Otra aproximación a la sinonimia

Además de una aproximación estructural a la sinonimia, cabe su tratamiento empírico. Confiados en que la labor del lexicógrafo recoge la competencia de los hablantes de un lenguaje natural sobre el uso flexible e informativo de las palabras, una tarea interesante es reflejar esta competencia con una semántica formal apropiada. Para ello debe sustituirse la pregunta de qué es la sinonimia por la de cómo medir la sinonimia. Las medidas de similaridad ofrecen un instrumento, bastante bien fundamentado y desarrollado, para comparar y cuantificar palabras que son sinónimas en diferente grado, y su traducción o implementación en una semántica de Prolog no es artificiosa.

Una vez que se dispone de un método automatizado para medir y tratar la sinonimia, se dispone de un medio eficaz para flexibilizar, de una manera automática, el lenguaje. Esto puede ser útil para un acceso más dúctil de los ordenadores a las bases de datos; para mejorar los interfaces de comunicación persona-máquina. La confección de un diccionario electrónico de sinónimos permite estudiar en detalle esto. Pero su estudio ha sido objeto de otro trabajo

(S. Fernández Lanza & A. Sobrino, 1998) y su exposición queda fuera del alcance de éste.

BIBLIOGRAFÍA

- Carnap, R. *Logical Foundations of Probability*, The University of Chicago Press, 1950.
- Fernández Lanza, S. & Sobrino, A. "Análisis del grado de sinonimia en un diccionario de sinónimos", *Actas del VIII Congreso Español sobre Tecnologías y Lógica Fuzzy*, 1998.
- Goodman, N. "On likeness of meaning", *Analysis*, 10, 1949.
- Lewis, C.I. "The modes of meaning", *Philosophy and Phenomenological Research*, vol 4(2), 1943.
- Mates, B. "Synonymy", Linsky, L. (ed) *Semantics and the Philosophy of Language*, University of Illinois Press, 1952.
- Naess, A. "Toward a theory of interpretation and preciseness", Linsky, L. (ed) *Semantics and the Philosophy of Language*, University of Illinois Press, 1952.
- Pelegrin, J., "Linguistics and Philosophy", *Theoretical Linguistics*, 24, 2-3, 1988, 245-264.
- Pereira, F. C. N. & Shieber, S. M. *Prolog and Natural Language Analysis*. CSLI. Center for the Study of Language and Information. Stanford, 1987.
- Quine, W.V.O. "Two dogmas of empiricism", *From a Logical Point of View*, Harvard University Press, 1953.
- Spark-Jones, K. *Synonymy and Semantic Classification*, Edinburg University Press, 1986.
- White, M. G. "The analytic and the synthetic: an untenable dualism", Linsky, L. (ed) *Semantics and the Philosophy of Language*, University of Illinois Press, 1952.

Santiago Fernández Lanza
Universidad de Santiago de Compostela
e-mail: sflanza@usc.es

Alejandro Sobrino Cerdeiriña
Universidad de Santiago de Compostela
e-mail: lflgalex@usc.es

LOS TÉRMINOS DE GÉNERO NATURAL COMO DESIGNADORES RÍGIDOS

LUIS FERNÁNDEZ MORENO
Universidad Complutense de Madrid

ABSTRACT: One of Kripke's main theses concerning natural kind terms is that they are rigid designators. Kripke does not explain, though, how he conceives the referents of natural kind terms. However without an elucidation of the referents of these terms the thesis that natural kind terms are rigid designators remains indefinite. The aim of this paper is to explicit and clarify the conception on the referents of natural kind terms that seems to underlie to that thesis of Kripke.

KEYWORDS: Natural kind, natural kind term, rigid designator, reference, Kripke.

Una de las tesis más importantes acerca de la semántica de los términos de género natural formulada por Saul Kripke en *Naming and Necessity* (Kripke [1972]), y mantenida igualmente en *Identity and Necessity* (Kripke [1971]), es que los términos de género natural son designadores rígidos. El objetivo de estas páginas es esclarecer dicha tesis y, en especial, los compromisos ontológicos que conlleva.

A este respecto conviene comenzar haciendo algunas observaciones acerca del marco en el que Kripke formula la tesis de que los términos de género natural son designadores rígidos. En primer lugar, Kripke no presenta una caracterización precisa de los términos de género natural, limitándose básicamente a indicar que los términos de género natural son términos generales y que entre ellos se encuentran términos contables, como "gato" o "tigre", y términos de masa, como "agua" y "oro";¹ si a esto unimos que Kripke emplea indistintamente las expresiones "términos de género natural" ("natural kind terms") y "términos de géneros naturales" ("terms for natural kinds"),² habría que concluir que los términos de género natural son caracterizados por Kripke simplemente como términos generales que designan - o, al menos, pretenden designar - *géneros naturales*, como los géneros gato o agua. Esta caracterización

¹ Kripke [1972], p. 134; véase también p. 127.

² Véase, p.e., Kripke [1972], p. 127.

sólo podrá adquirir perfiles más nítidos cuando se precise qué tipo de entidad son los géneros naturales o, si se prefiere, cómo se conciben los géneros naturales. En segundo lugar, es digno de mención que el tratamiento de los términos de género natural por parte de Kripke es secundario con respecto al de los nombres propios; sólo tras haberse ocupado de los nombres propios - en las dos primeras conferencias de [1972] - Kripke atiende - en la tercera y última conferencia de [1972] - a los términos de género natural, radicando su principal objetivo en señalar la existencia de *semejanzas* entre los términos de género natural y los nombres propios; justamente una de las afinidades existentes entre ambos tipos de términos consiste, según Kripke, en que ambos son *designadores rígidos* y dan lugar a *enunciados de identidad* que, en caso de ser verdaderos, son necesariamente verdaderos.

En la primera conferencia de [1972] Kripke introduce el término "designador" como denominación común de los nombres propios y de las descripciones definidas.³ La definición de la noción de designador rígido se encuentra en las conferencias primera y segunda de [1972] y, por tanto, aparece formulada antes de que Kripke atienda a los términos de género natural, mientras que la definición de dicha noción se presenta en [1971] cuando ya se han mencionado algunos ejemplos de términos de género natural, si bien en este escrito no se emplea la expresión "término de género natural" y, por tanto, no se afirma explícitamente que los términos de género natural son designadores rígidos. La definición de *designador rígido* que aparece en ambos escritos es la siguiente: un designador rígido es un término que designa el mismo objeto con respecto a todo mundo posible o, al menos, con respecto a todo mundo posible en el que exista dicho objeto.⁴ Puesto que en la tercera conferencia de [1972] Kripke extiende la noción de designador rígido a los términos de género natural, conviene sustituir la expresión "objeto", que aparece en dicha definición, por la expresión "entidad", de carácter más general, con lo que la caracterización kripkeana de designador rígido, aplicable tanto a términos singulares como a términos generales, sería la siguiente: un designador rígido es un término que designa la misma entidad con respecto a todo mundo posible o, al menos, con respecto a todo mundo posible en el que exista dicha entidad. Esta noción de designador rígido permite justificar la tesis de que los enunciados de identidad verdaderos en los que el signo de identidad aparezca flanqueado por dos designadores rígidos serán necesariamente verdaderos, es decir, verdaderos en todo mundo posible.⁵

³ Kripke [1972], p. 24.

⁴ Acerca de la caracterización más restringida de designador rígido véase Kripke [1972], pp. 48-49 y Kripke [1971], p. 79; sobre la definición más general véase Kripke [1971], p. 78 y [1972], pp. 48 y 77-78, así como p. 21, n. 21. En la caracterización de designador rígido presente en Kripke [1972], pp. 77-78 se emplea, en lugar de la expresión "objeto", la expresión "cosa".

⁵ De acuerdo con la definición más restringida de designador rígido habría que matizar esta tesis, indicando que los enunciados en cuestión serán verdaderos en todo mundo posible o, al

Al ocuparse de la semejanza mencionada entre los nombres propios y los términos de género natural, Kripke presta especial atención a un tipo de enunciados en los que aparecen términos de género natural y que poseen la forma de enunciados de identidad; él se refiere a tales enunciados como "identificaciones teóricas" o "identidades teóricas"⁶ y menciona como ejemplos los enunciados "El agua es H₂O" y "El oro es el elemento cuyo número atómico es 79". Kripke considera que una identidad teórica, como el enunciado "El agua es H₂O", es un enunciado de identidad que involucra dos designadores rígidos⁷ y es, por tanto, en caso de ser verdadero, necesariamente verdadero.⁸

Conviene señalar, sin embargo, que este rasgo que, según Kripke, comparten los nombres propios y los términos de género natural, a saber, el ser designadores rígidos y, consiguientemente, el dar lugar a enunciados de identidad que, en caso de ser verdaderos, son necesariamente verdaderos, es más fácilmente discernible con respecto a los primeros que con respecto a los segundos. Esto es así, principalmente, porque Kripke identifica, como es usual, el referente o entidad designada por un nombre propio -digamos su *designatum*-,

menos, en todo mundo posible en el que existan las entidades designadas por dichos términos. En lo siguiente pasaré por alto esta puntualización.

⁶ Kripke emplea generalmente la primera expresión, pero, al menos en una ocasión (Kripke [1972], p. 140), utiliza la expresión "identidad teórica"; considero preferible esta segunda expresión, de la que me serviré preferentemente a partir de ahora. Conviene añadir que de algunas observaciones de Kripke (p.e., en Kripke [1972], p. 116) se seguiría que las identificaciones teóricas o identidades teóricas no son tanto enunciados de identidad cuanto lo expresado - las proposiciones expresadas - por dichos enunciados; puesto que esta cuestión no afecta a mi objetivo en estas páginas, y en aras de facilitar mi exposición, voy a interpretar las identidades teóricas en el primer sentido, es decir, como un tipo de enunciados de identidad.

⁷ En una ocasión Kripke parece insinuar la existencia de excepciones a esta tesis, al afirmar que "[l]as identidades teóricas [...] son *generalmente* identidades que involucran dos designadores rígidos" ([1972], p. 140; la cursiva es mía), pero Kripke no menciona, al menos explícitamente, ejemplos de dichas excepciones. En lo siguiente no tomaré en cuenta esta matización, pues no parece aplicarse a las identidades teóricas paradigmáticas, como las de los ejemplos mencionados.

⁸ Cabría mantener que no todas las identidades teóricas conciernen términos de género natural, alegando que cabe distinguir los términos de género natural en sentido estricto del tipo de términos a los que Kripke se refiere como "términos de fenómenos naturales", entre los que se encuentran los términos "calor" y "luz". Algunas afirmaciones de Kripke parecen indicar claramente que él subsume los términos de fenómenos naturales bajo los términos de género natural (véase, p.e., Kripke [1972], p. 134), si bien otras afirmaciones suyas podrían servir de base para mantener que Kripke distingue los fenómenos naturales de los géneros naturales - véase, p.e., [1972], p. 123, n. 63 y p. 138 -, lo que permitiría distinguir entre ambos tipos de términos, por cuanto designarían distintos tipos de entidades - pero véase nota 21 *infra*. No obstante, ésta sería la única diferencia semántica importante entre ambos tipos de términos que cabría atribuir a Kripke, pues, al igual que ocurre con los que cabría considerar como términos de género natural en sentido estricto, Kripke mantiene que los términos de fenómenos naturales son designadores rígidos, aparecen en identidades teóricas, ... Puesto que, en cualquier caso, la mayoría de las afirmaciones de Kripke acerca del primer tipo de términos se aplican también al segundo - y viceversa -, en lo siguiente pasaré por alto tal posible distinción.

con el objeto o individuo designado por el nombre, pero Kripke dista de aclararnos cómo concibe el referente o *designatum* de un término de género natural. En los contextos relevantes Kripke se sirve de expresiones como "sustancia", "especie", "propiedad" o "género" sin ninguna clarificación, usando a menudo algunas de estas expresiones como intercambiables.⁹ Por una parte, esto no es de extrañar, pues tanto sustancias como especies serían tipos de géneros; por otra parte, empero, es inusual emplear de manera indistinta, como en ocasiones hace Kripke, las expresiones "género" y "propiedad", aunque posteriormente propondré una interpretación al respecto. En cualquier caso, sin una mínima precisión acerca de los referentes o *designata* de los términos de género natural la tesis de que estos términos son designadores rígidos será igualmente imprecisa; dicho de otra manera, a la pregunta de cuáles son los referentes de los términos de género natural cabe dar una respuesta trivial, a saber, son géneros o, más bien, géneros naturales, pero mientras no se especifique cómo se conciben tales entidades la tesis de que los términos de género natural son designadores rígidos y dan lugar - o, al menos, pueden dar lugar - a enunciados de identidad necesariamente verdaderos quedará sumida en la oscuridad. Más aún, el que los términos de género natural sean o no designadores rígidos dependerá de cómo se conciban los géneros naturales; por esta razón mi objetivo consistirá en encontrar una concepción de los géneros naturales que concuerde, al menos parcialmente, con afirmaciones de Kripke y que permita que los términos de género natural sean designadores rígidos.

Como Kripke señala, los términos de género natural son de dos tipos, términos de masa y términos contables. Ahora bien, conviene señalar que los términos de masa pueden considerarse no sólo como términos generales, sino también como *términos singulares* y que esta interpretación puede extenderse a los términos contables, lo que permitiría considerar a todos los términos de género natural como términos singulares. Esta interpretación de los términos de género natural no debe pasarse por alto, especialmente si se hace hincapié en las identidades teóricas, pues el signo de identidad en su empleo más básico, es decir, en el ámbito de la lógica de primer orden, vincula dos términos singulares.

Si los términos de género natural se consideran como términos singulares y se pretende mantener que sus referentes - es decir, los géneros naturales - son *entidades concretas*, el marco más adecuado de interpretación consiste en la adopción de un *enfoque mereológico*, con lo que los referentes de los términos de género natural serían totalidades mereológicas, que son concebidas como individuos que constan de diversas partes; puesto que estas partes serían entidades concretas - por ejemplo, tigres o porciones de agua -, la totalidad

⁹ Acerca del empleo indistinto de los términos "sustancia" y "género" véase, p.e., Kripke [1972], pp. 119 y 136; sobre el uso indistinto de los términos "especie" y "género", véase, p.e., Kripke [1972], pp. 121 y 127. Por lo que concierne al empleo de los términos "propiedad" y "género" véanse las notas 18 y 19 *infra*.

mereológica correspondiente sería asimismo una entidad concreta. No obstante, a este respecto conviene señalar que - aparte de que en los escritos donde Kripke se ocupa de los términos de género natural no hay alusión alguna a una interpretación mereológica de los referentes de estos términos¹⁰ - la mayoría de los enfoques mereológicos adoptan el principio de extensionalidad mereológica, según el cual la identidad de una totalidad viene determinada por sus partes, pero, puesto que las entidades que constituirían estas partes podrán ser diferentes en distintos mundos posibles, la totalidad mereológica en cuestión no será invariante con respecto a distintos mundos posibles, a consecuencia de lo cual los términos de género natural así interpretados no podrían ser designadores rígidos.

De acuerdo con lo anterior, si se pretende unificar la interpretación de los términos de género natural, ya sean términos de masa o términos contables, al mismo tiempo que se los considera como términos singulares y como designadores rígidos, la opción más plausible es mantener que sus *designata* son, no entidades concretas, sino *entidades abstractas* - ya las denominemos "sustancias", "especies" o "géneros". Un representante de esta posición es K. Donnellan, quien estima que, si se pretende mantener que los términos de género natural son designadores rígidos, la propuesta más recomendable consiste en interpretarlos como "nombres abstractos", es decir, como términos singulares que designan entidades abstractas; así Donnellan afirma acerca del término "tigre" que "sólo se convierte en un designador rígido cuando es considerado como un nombre abstracto que denota la tigriedad o la especie tigre".¹¹ Donnellan no especifica cómo concibe la relación entre tales entidades abstractas y las entidades concretas - por ejemplo, entre la tigriedad y los tigres -; si suponemos que esta relación es la de instanciación, habría que afirmar que dichas entidades abstractas estarán instanciadas - o, al menos, podrán estarlo - en entidades concretas. Y llegados a este punto cabe señalar que algunas afirmaciones de Kripke concuerdan, efectivamente, con la concepción de los *designata* de los términos de género natural como entidades abstractas instanciadas en entidades concretas; así, por ejemplo, Kripke afirma:

"Creo que, en general, la referencia de los términos de géneros naturales (p.e., de géneros animales, vegetales y químicos) se fija de esta manera: la sustancia es definida como el género *instanciado* por (casi la totalidad) de una muestra dada".¹²

¹⁰ Si hay, en cambio, una alusión a una concepción mereológica de los mundos posibles o, más precisamente, de los estados posibles del mundo, de la que Kripke discrepa; véase Kripke [1972], pp. 16 s.

¹¹ Donnellan [1973], p. 712.

¹² Kripke [1972], pp. 135 s.; la cursiva es mía. La observación entre paréntesis se debe, obviamente, a que entre las entidades en base a las cuales se fija la referencia de un término de

Si los términos de género natural se interpretan de la manera indicada, es decir, como términos singulares que designan entidades abstractas, los términos de género natural serán designadores rígidos, pues las entidades abstractas que constituyen sus *designata* son invariantes con respecto a todo mundo posible o, al menos, con respecto a todo mundo posible en el que se encuentren instanciadas; asimismo, de acuerdo con dicha interpretación, los términos de género natural podrán dar lugar a enunciados de identidad necesariamente verdaderos. Ahora bien, si los términos de género natural se interpretan como términos singulares que designan entidades abstractas, existiría una semejanza obvia entre ellos y los nombres propios, a saber, ambos tipos de términos serían *términos singulares*. Ésta es una semejanza que Kripke no menciona, pero esto no es de extrañar, pues él afirma que los términos de género natural son *términos generales*, no términos singulares. Por este motivo conviene que pasemos a considerar a los términos de género natural como términos generales.

Por el referente de un término general suele entenderse su *extensión*, es decir, el conjunto de individuos a los que el término se aplica. Ahora bien, si se afirma que los términos de género natural, como el término "gato", son designadores rígidos, su referente no podrá ser su extensión, es decir, un conjunto - en el caso del término "gato", el conjunto de los gatos -, pues este conjunto podrá tener diferentes elementos en distintos mundos posibles; es decir, la extensión de un término general no es la misma en todos los mundos posibles. Por tanto, si los términos de género natural son, como pretende Kripke, términos generales y designadores rígidos, sus referentes no podrán ser sus extensiones.¹³

Hay autores que, teniendo en cuenta que si los términos de género natural son términos generales y designadores rígidos, no pueden tener como referentes sus extensiones, proponen que tengan como referentes, no sus extensiones, sino los *miembros* de dichas extensiones. Uno de estos autores, M. Cook, tomando como ejemplo el término "gato", alega que el referente del término "gato" en un mundo posible no es el conjunto de los gatos existentes en ese mundo posible, es decir, la extensión del término "gato" en dicho mundo posible, sino los gatos existentes en ese mundo posible, es decir, todos y cada uno de los gatos existentes en dicho mundo posible, pero, en opinión de Cook, esto no es óbice para que el término "gato" sea un designador rígido; según Cook, dicho término es un designador rígido por cuanto, "aunque no designa los mismos objetos en

género natural pueden encontrarse algunas entidades de un género distinto; a este respecto hablamos en ocasiones de "impurezas".

¹³ Una manera de evitar esta dificultad consistiría en considerar como referente de un término de género natural no su extensión en cada mundo posible, sino la unión de dichas extensiones, entendiendo por tal un conjunto de individuos actuales y posibles; por ejemplo, el referente del término "gato" sería el conjunto de los gatos actuales y posibles. Pero Kripke no aceptaría tal concepción de los referentes de los términos de género natural, por cuanto no comparte el realismo modal que está a la base de dicha concepción.

todos los mundos posibles, designa los mismos objetos *dondequiera que esos objetos existan*".¹⁴

Dos aspectos de la posición de Cook son dignos de mención. Por una parte, dicha posición presupone que una propiedad esencial de cada gato es la de ser un gato, por lo que un individuo que en un mundo posible sea un gato ha de serlo en todos los mundos posibles en los que dicho individuo exista y, por tanto, en todos esos mundos posibles será designado mediante el término "gato"; dicho de otra manera, la posición de Cook presupone que la pertenencia de un individuo a un género es un rasgo esencial de dicho individuo. Éste es un aspecto en el que la posición de Cook puede concordar con la de Kripke, pues algunas observaciones de Kripke apuntan a que él también mantiene el esencialismo por lo que respecta a la pertenencia a los géneros.¹⁵ Por otra parte, de acuerdo con la posición de Cook la referencia de los términos de género natural no es su referencia usual - su extensión - sino una referencia distributiva: el término "gato" "designa todos y cada uno de los gatos, dondequiera que exista(n)".¹⁶ Ahora bien, a este respecto cabe preguntarse si es posible mantener al mismo tiempo que la referencia de los términos de género natural es una referencia distributiva y que estos términos cumplen la definición de designador rígido formulada por Kripke.

Cook toma como base de sus consideraciones la caracterización kripkeana de la noción de designador rígido según la cual un designador rígido es un término que designa la misma entidad con respecto a todo mundo posible en el que exista dicha entidad. Pero, como, según Cook, el *designatum* de un término de género natural, como el término "gato", es múltiple, a saber, todos y cada uno de los gatos, la caracterización de designador rígido que Cook ha de adoptar *difiere* de la de Kripke y sería la siguiente: un designador rígido es un término que designa la misma entidad o *las mismas entidades* con respecto a todo mundo posible en el que exista dicha entidad o en el que existan *dichas entidades*.¹⁷ No obstante, cabría alegar que la diferencia existente entre ambas caracterizaciones de la noción de designador rígido podría deberse a que Kripke habría formulado su definición de designador rígido exclusivamente para términos singulares, pero que, de haberla formulado para términos singulares y términos generales, habría propuesto una definición semejante a la que subyace a la posición de Cook.

En cualquier caso, hay otra razón por la que Kripke no habría aceptado que la referencia de los términos de género natural fuese, como para Cook, una

¹⁴ Cook [1980], p. 63.

¹⁵ Véase Kripke [1972], p. 115, n. 57, así como p. 135, n. 69.

¹⁶ Cook [1980], p. 63.

¹⁷ Esta caracterización de la noción de designador rígido sería aplicable tanto a términos singulares como a términos generales, pues, de acuerdo con la posición de Cook, la caracterización de designador rígido para los términos generales sería la siguiente: un designador rígido es un término que designa las mismas entidades con respecto a todo mundo posible en el que existan dichas entidades.

referencia distributiva, pues, en ese caso, los enunciados que Kripke denomina "identidades teóricas" no tendrían la forma de enunciados de identidad, sino la de cuantificaciones universales de bicondicionales, estando constituidos los miembros de estos bicondicionales por oraciones abiertas formadas con las constantes de predicado mediante las que simbolizaríamos los términos generales en cuestión. Así un enunciado como "El agua es H_2O " sería parafraseado como afirmando que para todo individuo (del universo de discurso) se cumple que ese individuo es (una porción de) agua si y sólo si ese individuo es (una porción - compuesta de moléculas - de) H_2O y sería formalizado, por ejemplo, como " $\forall x (Px \leftrightarrow Qx)$ ". Es decir, si se considera que los *designata* de los términos de género natural son en cada mundo posible (todos y cada uno de) los miembros de sus extensiones - en ese mundo posible -, los enunciados que Kripke denomina "identidades teóricas" y que él considera que poseen la forma de enunciados de identidad *no* poseerían la forma de enunciados de identidad. Ésta parece ser una razón suficiente para excluir que Kripke hubiera aceptado la interpretación de los términos de género natural como designadores rígidos propuesta por Cook.

Habiendo admitido que si los términos de género natural son términos generales y designadores rígidos, sus *designata* no pueden ser ni sus extensiones ni los miembros de dichas extensiones, cabría plantear la cuestión de si las entidades designadas por dichos términos no serán quizás sus *intensiones*, que con frecuencia han sido denominadas "propiedades". Una razón adicional para traer a colación en este contexto la alusión a propiedades radica en que algunas observaciones de Kripke en [1972] apuntan claramente a que él considera que las identidades teóricas expresan una identidad entre *propiedades*,¹⁸ si bien estas afirmaciones de Kripke pierden gran parte de su fuerza por el hecho de que él afirma explícitamente que está empleando el término "propiedad" en un sentido lato, que, lamentablemente, no explicita.¹⁹ En cualquier caso, puesto que, de acuerdo con una concepción acerca de las propiedades, éstas son las intensiones de expresiones lingüísticas como los términos generales - o los predicados correspondientes -, cabe plantear la cuestión de si, tal como Kripke concibe las entidades designadas por los términos de género natural, éstas son las intensiones de dichos términos.

¹⁸ Kripke [1972], pp. 138 y 139; véase también p. 140, n. 71. Estas observaciones de Kripke sugieren la equiparación de géneros y propiedades, lo que justificaría el empleo indistinto de los términos "género" y "propiedad". En [1972], p. 148, Kripke se refiere a ciertas identidades teóricas como si expresasen una identidad de tipos, pero esto es compatible con lo anterior, pues en algunas ocasiones Kripke parece emplear indistintamente las expresiones "tipo" y "propiedad"; véase a este respecto, p.e., [1972] p. 144 y nota 73 de esa misma página.

¹⁹ Kripke afirma en [1972], p. 137: " [El término] 'p[ropiedades]' es usado aquí en un sentido lato y puede incluir géneros más amplios; por ejemplo, animalidad y felinidad, para los tigres". Este texto constituye uno de los ejemplos en los que Kripke emplea con cierta indistinción los términos "género" y "propiedad".

Ahora bien, al menos de acuerdo con dos concepciones posibles acerca de las intensiones de expresiones lingüísticas, la respuesta a esta cuestión ha de ser negativa. De acuerdo con una de estas concepciones, las intensiones son funciones cuyos argumentos son mundos posibles y cuyos valores son extensiones, pero puesto que las extensiones de los términos de género natural no son invariantes con respecto a los mundos posibles, tampoco lo serán sus intensiones y, por ende, las propiedades así entendidas. De acuerdo con otra concepción acerca de las intensiones, que puede servir de apoyo a la anterior, la intensión de un término se identifica con su sentido o significado, el cual se concibe como proporcionando condiciones necesarias y suficientes para la aplicación del término. Pero, aunque las propiedades así entendidas sí poseen la invariancia con respecto a los mundos posibles exigida del *designatum* de un designador rígido, esta interpretación de los *designata* de los términos de género natural no concuerda con el carácter de las identidades teóricas, pues éstas no afirman la identidad de significado o sinonimia de dos expresiones, sino que pretenden expresar descubrimientos científicos; en la terminología preferida por Kripke, las identidades teóricas son enunciados necesarios *a posteriori*, no enunciados analíticos, es decir, no enunciados necesarios *a priori*.

No obstante, la concepción acerca de las propiedades que hemos mencionado, según la cual éstas son intensiones de expresiones lingüísticas, no es la única posible. Efectivamente, hay otra concepción más adecuada para nuestros fines, según la cual las propiedades son entidades integrantes de nuestro mundo y descubribles mediante investigación empírica. Las propiedades son concebidas como entidades abstractas, si bien se hallan instanciadas en entidades concretas - recuérdese el texto de la nota 12 - o son constituyentes de entidades concretas, y son responsables de las relaciones causales entre éstas. En este sentido el enunciado "El agua es H₂O" expresaría una identidad entre propiedades, a saber, entre la propiedad de ser agua y la propiedad de estar compuesto de moléculas de H₂O o, dicho de manera más abreviada, la propiedad de ser H₂O. Si los términos de género natural se interpretan de esta manera, es decir, como términos generales que designan propiedades,²⁰ dichos términos poseerán el carácter de designadores rígidos, por cuanto designarán la misma propiedad con respecto a todo mundo posible o, al menos, con respecto a todo mundo posible en el que se encuentre instanciada la propiedad en cuestión.²¹

²⁰ Para distinguir esta segunda concepción de las propiedades de la primera, según la cual las propiedades eran intensiones de expresiones lingüísticas, podríamos referirnos al segundo tipo de propiedades como propiedades empíricas.

²¹ Llegados a este punto cabría plantear la cuestión acerca de cuáles son los *designata* de los términos que Kripke denomina "términos de fenómenos naturales" - véase nota 8 *supra*. Si consideramos que estos términos quedan subsumidos bajo los términos de género natural, habría que responder que sus referentes son propiedades; si, por el contrario, consideramos que no son términos de género natural - aunque estén emparentados con ellos - podríamos responder que sus referentes son *tipos* de fenómenos naturales - véase Kripke [1972], p. 148 -, aunque, a tenor

Ahora bien, puesto que los términos de género natural están siendo considerados como términos generales y éstos serán simbolizados mediante signos de predicado, las identidades teóricas correspondientes serán enunciados de identidad formulables en el lenguaje de la lógica de segundo orden; así el enunciado "El agua es H_2O " podría ser formalizado como " $P=Q$ ".

De acuerdo con lo anterior, la interpretación más plausible de las afirmaciones de Kripke acerca de los términos de género natural y, en concreto, de sus afirmaciones acerca de que dichos términos son términos generales, son designadores rígidos y dan lugar a enunciados de identidad que, en caso de ser verdaderos, son necesariamente verdaderos, pero que son necesarios *a posteriori*, no *a priori*, es considerar que los referentes de los términos de género natural y, por tanto, los géneros naturales son propiedades, concebidas como entidades abstractas, pero instanciadas en entidades concretas y causalmente operantes. Por esta razón no debe extrañarnos que, como he indicado, Kripke emplee en ocasiones de manera indistinta los términos "género" y "propiedad".

BIBLIOGRAFÍA

- Cook, M.: "If 'Cat' Is a Rigid Designator, What Does It Designate?". *Philosophical Studies*, 37, 1980, pp. 61-64.
- Donnellan, K.: "Substances as Individuals". *Journal of Philosophy*, 70, 1973, pp. 711-712.
- Kripke, S.: "Identity and Necessity", en M. Munitz (ed.), *Identity and Individuation*, Nueva York, New York University Press, 1971, pp. 135-164. Reimp. en S. Schwartz (ed.), *Naming, Necessity and Natural Kinds*, Ithaca, Cornell University Press, 1977, pp. 66-101.
- Kripke, S.: "Naming and Necessity". En D. Davidson y G. Harman (eds.), *Semantics of Natural Languages*, Dordrecht, Reidel, 1972, pp. 253-355 y 763-769. Edición revisada y ampliada: *Naming and Necessity*, Oxford, Blackwell, 1980.

Luis Fernández Moreno

Universidad Complutense de Madrid

e-mail: luisfm@eucmos.sim.ucm.es

de lo indicado en la nota 18, estas afirmaciones pueden ser equivalentes; en cualquier caso, los referentes de los "términos de fenómenos naturales" serían entidades abstractas.

TOWARDS A REFORMULATION OF THE ANALYTICAL CONCEPT OF THE HISTORY OF PHILOSOPHY

HENRIQUE JALES RIBEIRO
University of Coimbra

ABSTRACT: The author discusses some important historical and philosophical presuppositions concerning the analytical concept of the history of philosophy, suggesting, finally, the need for a reformulation of this concept. The concept of "analytical tradition", like the concept of "origins of analytical philosophy", is clearly something which, today, should be understood historically, in contrast with the metahistorical and metaphilosophical use of the past, that is to say, not in general or abstract terms but in relation to each particular historical and philosophical context in the development of analytical philosophy. From this perspective, it is suggested that we must dispense with a series of presuppositions that are more or less fundamental to analytical historiography and to what is supposed to be analytical philosophy itself. The need for a reformulation of the analytical concept of history of philosophy necessarily leads, apparently at least, to a reformulation of analytical philosophy itself.

KEYWORDS: Analytical philosophy, history of analytical philosophy, origins of analytical philosophy, contexts of analytical philosophy, presuppositions of analytical philosophy.

The historian of philosophy, faced with a series of historical and philosophical problems to explain, frequently find himself in a similar situation to the man of science, in that he suspects that current theories do not adequately explain the problems. This may be either because they are only valid in part, leaving the rest unexplained, or because they only manage to explain these problems in general terms, leaving out material which in fact constitutes the basis of the respective interpretation. Like the theoretical physicist or mathematician, the historian of philosophy basically has two options in this situation: he may continue to use these theories, inevitably sacrificing the possibility of greater elucidation, or instead, he may consciously and deliberately seek out another theory, or theories, which allow him to achieve objectives that are beyond the range of the others. In this second situation, it may not be necessary to reject the inadequate or inappropriate

theories completely, because his contribution may in fact take the form of a reformulation and extension of what went before, as frequently happens in physics and the other natural sciences. In any case, he may not necessarily be aware of this when initiating his studies with the presupposition that a certain given theory is inadequate or inappropriate. Moreover, even when a new theory arises during the course of his investigations that implies the rejection of previously established theories, the historian of philosophy constantly needs to compare it with others in order to demonstrate when, how and where it is more competent. This might be either because it explains certain problems that other theories cannot, whilst simultaneously throwing new light on others that were explained in some way, or because it brings to the question a new and unexpected set of basic questions that were far from being understood previously.

This outline of research procedure in the history of philosophy may seem a simplistic caricature, based upon analogy with the natural sciences. However, it is clear that the historian of philosophy, like the theoretical physicist or any other scientist, is never confronted with some virgin territory of "facts" constituting the documental history of philosophical thought. These facts are always interpreted or mediated by previously established theories. Indeed, we need to understand that there is no natural kind of historical and philosophical fact that permanently exists in history awaiting our commentary; and, of course, even if such a body of facts did exist, it would certainly have been discovered and thoroughly analysed a long time ago, thus making redundant any kind of investigation. On the contrary, historical and philosophical facts that are supposedly neutral and free from interpretation are already permeated by theory when our attention is first drawn to them; they are saturated in current theories which form part of the historical and philosophical context of interpretation into which the research is inserted. The question, therefore, is not about setting aside the theories to gain access to the facts, but (since, in spite of everything, facts do exist, as the vast amount of documentation about a philosophical problem proves, about showing whether the theories adequately explain these facts, and if not, proposing new and more competent ones. Obviously, it is not possible to escape this basic hermeneutical circle between theory and facts; it may be a temptation to assert either the traditional position that only "facts" exist, or the more recent corollary that only "theories" exist, but both of these mean renouncing the basic methodological precepts of scientific research, which takes as its starting-point some aspect which awakens the need for research (a group of problems that have not been properly understood or satisfactorily explained),

the interpretation of which requires a new theory which, when understood, will eventually become object of criticism and discussion in its own right.

These methodological considerations provide the justification for the objectives of my paper: the problems raised by the interpretation of the history of analytical philosophy. One of the most prominent and current features of these problems is its metahistorical and metaphilosophical character. This means that the problems in question about Frege or Russell or Moore, for example, do not arise directly from their philosophies in themselves; that is to say, they are not necessarily problems raised expressly or implicitly by these philosophers, but rather are problems of interpretation arising from the historical and philosophical contexts in which analytical philosophy developed. In other words, they are problems, in general, of certain more recent historical and philosophical contexts, and not philosophical problems posed by Frege or Russell or Moore, either directly or indirectly. This does not mean that Russell himself, for instance, did not ever attempt to respond to metaphilosophical questions; on the contrary, these are so inseparable from his philosophical reflection that it is sometimes difficult or even impossible to disentangle them. It does mean, however, as previously stated, that they relate more to more recent philosophical contexts, which, in general, are different from the original contexts of those philosophers. Examples of the first type of problems are: was Russell (or Frege or Moore) the founder of analytical philosophy? what is the predominant trait in the conception of logical analysis which gave rise to the analytical movement, and to what extent is this conception in accordance with present conceptions? what is Russell's (or Frege's or Moore's) contribution to the idea that the philosophy of language constitutes, in fact, the foundations of philosophy in general?

Let me take Russell's philosophy as a case study. The fact that Russell himself did not raise these types of questions clearly does not mean that they are not pertinent, maybe even historically and philosophically fundamental. Whatever questions we ask of philosophers who are absent from the present historical and philosophical context, they will always be essentially contextual questions. This is not a debatable philosophical thesis but a lesson of philosophical historiography in general. With regards to interpretations of Russell, the commentaries of L. Stebbing or J. Wisdom in the mid-thirties are naturally different from those of the English ordinary language philosophers in the forties and fifties (M. Black, P. Strawson, G. Ryle, etc.); and these are different again from more recent interpretations, such as those of R. Rorty and M. Dummett. A particularly rich and far-reaching example, which continues

today to stimulate research, is the case of the commentaries made in the twenties by some of the authors behind the Vienna Circle (M. Schlich, O. Neurath, P. Frank or R. Carnap): these commentaries are generally different from those made by the same authors at the end of the decade when the Circle was formally set up, and different again from their comments in the mid-thirties when, some years before its dissolution, the group was at the height of its glory. All the cases listed here have one point in common which explains the differences of interpretation: the historical contexts which serve as a backdrop to these different commentaries are themselves philosophically different. It is not enough to assert that they are contexts of "analytical philosophy" or of the "analytical movement", since these are clear examples of metahistorical and metaphilosophical concepts originating in a more recent historical and philosophical context. In fact, as regards "analytical philosophy", it is particularly important to identify the conjunction of historical and philosophical circumstances underpinning these commentaries, thereby defining an "interpretation context", because what we understand today as "analytical philosophy" seems to be a historical result of not only the development of philosophical ideas in general, but also of ideas that were originally metahistorical and metaphilosophical.

These are important reasons for us to doubt the ability to explain and therefore the legitimacy of the current analytical concept of the history of philosophy. As I have suggested before, the concept of "analytical tradition" is clearly something which today should be understood historically, that is to say, within its respective emergent context. It is important to look carefully at how and why the analytical movement managed to arrive at the position of being able to speak of itself as constituting a special philosophical tradition distinct from the (supposed) "continental tradition". It is also important to examine why this happened at a certain time and not at any other, and what type of justification it was supposed to provide for the philosophical reflection produced within that context. From the point of view of the former, it is possible to show that this concept is essentially a metahistorical source of legitimacy for analytical practice, for example, for the practice of English ordinary language philosophy, and that it did not initially have the universality that was frequently supposed, due to the fact that it excluded logical positivism from the field of analytical philosophy, if not in its form, at least in its intentions. But the main objection to the historical and philosophical legitimacy of this concept originates in the specialized research done today into the origins of the analytical movement, despite the fact that some of this is undertaken under the banner of the "analytical tradition" or its substitutes.

The studies made by R. Haller, A. Coffa and M. Friedman, amongst others, clearly show the essential dependence of contemporary analytical philosophy in general on Viennese logical positivism, and its origins in the continental philosophical tradition from which the analytical tradition is supposed to have emerged, not as an successor but as an alternative and opposing movement. The so-called "analytical tradition" appears in this light to be more of a metahistorical justification of a particular concept of historically-situated analysis (which requires its own context to be duly investigated), rather than a historical reality in itself or a methodological presupposition that should orientate our investigations into the history of analytical philosophy. Some studies perceive in Russell's philosophy an expression of perspectives that are alien to pure analytical philosophy and in fact reveal the influence of the Continental Cartesian tradition—an argument that would seem particularly pertinent in historical and philosophical terms, though not for the reasons that are presented. It is not a question of comparing Russell's philosophy with the analytical tradition, and of integrating it within it or dissociating it from it, but of understanding, finally, that two such distinct and historically divorced philosophical traditions do not exist.

Very similar considerations may be made about the question of the "origins of analytical philosophy" which is at the root of the concept of "analytical tradition". This is, again, a metahistorical and metaphilosophical question which should not be couched in general or abstract terms but in relation to each particular historical and philosophical context in the development of the analytical movement. This does not mean that it is not an important question, capable of directing our research to a certain extent; it does, however, mean that it is far from having the historical and philosophical reach that is usually claimed. Whether analytical philosophy began with Russell or Moore or Wittgenstein is a problem that cannot be decided by historical research, because the presuppositions of the problem are mainly our own general presuppositions and not the presuppositions of the philosophers in question. Without abandoning historical research, it is possible to equate and understand such a problem and reformulate it in terms of what we could call a "history of the history of analytical philosophy" (that is, a history of the concept of the history of analytical philosophy). This point of view may demonstrate first and foremost that historically there is not one single origin of analytical philosophy, but that each historical and philosophical context has metahistorically conceived its own historical origin; there is therefore no origin of analytical philosophy as such, but as many origins as the concepts of origin produced by the diverse contexts of the evolution of the analytical movement.

From this perspective it is possible to demonstrate that the concept of the "origins of analytical philosophy" was introduced essentially as a standard source of justification for analytical practice in each of these contexts; the representation of the historical period surrounding its "origins" was idealised and became perceived as a legitimate standard of "pure analysis" against which all applications were measured. With reference to Russell's philosophy, for example, it is suggested throughout my doctoral thesis that the use of such a concept for the investigation of problems which characterize it is methodologically wrong and philosophically misleading, and that this research should pass without the metahistorical and metaphilosophical arsenal of modern researchers.

In dispensing with a series of presuppositions that are more or less fundamental to contemporary philosophical historiography, we are not simply casting aside methodological orientations. These are essentially the philosophical presuppositions of what is supposed to be analytical philosophy itself: it is, therefore, this last type of presupposition that may generally be dispensed with. That analytical philosophy makes the philosophy of language the foundation of the rest of philosophy, that it rejects epistemology and ontology as the source of real philosophical problems, or identifies them as language problems—these are analytical criteria of interpretation that are generally shared and need to be abandoned as far as possible when we are dealing with philosophers such as Frege or Russell or Moore. But the same reasons that led us to abandon these types of presuppositions also indicate the possibility of a reinterpretation of the various context-bound interpretations of the place of these philosophers in the history of analytical philosophy. Concerning Russell for example, the idea that his philosophy has led to the reductionism of logical positivism in general, as Quine argued, that it is essentially alien to behaviourist naturalism and to the theory of meaning with basis in use, as Putnam asserted, or that it has been essentially unable throughout its history to understand and accept new and more modern perspectives proposed by Wittgenstein, as P. M. S. Hacker alleges, are questionable metahistorical and metaphilosophical presuppositions that we must understand within their own contexts as standard justifications of their respective conceptions of analysis.

This last point would seem to be fundamental. Contemporary philosophical history has suggested for a long time that not only were the analytical histories of the great authors inseparable from their contributions to philosophy, but also that they were perfectly valid and faithful presuppositions of historical research. If Popper saw in the philosophy of

Carnap the expression of an ingenuous reductionism and naturalism which characterises in general Logical Positivism's theory of meaning and which could have its origins directly in the philosophy of Russell, and if systematic criticism of these aspects were one of the basic presuppositions of the emergence of Popper's philosophy, then it would appear incontrovertible they constituted a historical reality. However, in fact there is no reason to think that Popper's or anyone else's historical analysis is inseparable from his philosophical contribution to such an extent that the one without the other would lose its originality and philosophical relevance. There seems to be no doubt that one of the presuppositions of Popper's philosophy, as early as 1934, was effectively that one ; but this does not in any way suggests that such a presupposition has a true historical correlate, and, therefore, that Popper's interpretation was historically correct. In the same way, when Quine (at more or less the same time as Popper) saw an equally ingenuous reductionism and naturalism in Russell's theory of meaning, which could be at the origins of logical positivism and which contrasted sharply with the naturalist and anti-foundationalism holism of his own philosophy, this is clearly a metahistorical and metaphilosophical representation of analytical history carried out in a debatable analytical interpretation context.

Finally, I would like to suggest that if contemporary analytical philosophy is in crisis, as we repeatedly hear, perhaps the reformulation of the concept of its own history, as suggested in this paper, could provide part of the solution, because in a sense it may be said that this crisis is just a crisis of the concept in question.

Henrique Jales Ribeiro
University of Coimbra
e-mail: jales@ci.uc.pt



NAMING AND COORDINATION

JOSEP MACIÀ FÀBREGA
University of Barcelona

ABSTRACT: Saul Kripke suggests in *Naming and Necessity* that descriptivist theories of proper names are not compatible with the very plausible general picture about how proper names work that he presents. We will claim that for speaker B to understand speaker's A use of a name N it is required that A and B are, as we will say, *coordinated* in their use of N. This coordination requirement ensures that the Kripkean picture obtains even if a descriptivist theory is correct. The coordination requirement actually holds of all expressions, including "bachelor" or "fortnight" for which there is no question that they have some associated descriptive material.

KEYWORDS: Names, descriptions, meaning, conventions.

1. Introduction

One of the main aims of Naming and Necessity (N&N) is to show that it is very implausible that any descriptivist theory of proper names be correct. Another aim is to present an alternative picture of how proper names work. The aim of this paper is to show that a certain version of descriptivist theory is perfectly compatible with what we will call *the Kripkean picture of proper names*; in fact we will see that the specific descriptivist theory we will present *requires* the truth of the Kripkean picture.

According to Kripke's characterization in N&N, a descriptivist theory of proper names (DT) holds that a name has an associated description or cluster of descriptions, and that if an object fits the description or most of the descriptions in the cluster, the name refers to that object, and if nothing fits the description the name does not refer. What we will call a *meaning descriptivist theory* holds that the associated description is synonymous with the name; what we will call a *reference-fixing descriptivist theory* holds that the associated description only fixes the referent of the name.

Kripke acknowledges that there are some very powerful reasons in favour of a descriptive theory of proper names. One of them is that a descriptivist theory can explain how the reference of a proper name is

determined, whereas for a non-descriptivist theory it would seem mysterious how a name can *reach* its referent. "If there is *not* such a descriptive content to the name, then how do people ever use names to refer to things at all?" [p. 28]¹. There is a deeply felt intuition that in order to be speaking about something by using a name one must have, as Wettstein(1989) puts it, a "*cognitive fix* on the thing in question", "something in one's thought must correctly distinguish the relevant item from everything else in the universe"².

The advantages of descriptivism notwithstanding, Kripke holds that descriptivism regarding proper names is clearly false. He presents a number of very serious difficulties for a DT:

(I) The modal argument: if we substitute any of the most plausible candidates to be the description associated with "Aristotle" for x in 'Aristotle might not have been the x' we obtain a true statement.

(II) Non-necessity (of the associated description): In the Godel-example, "Godel" will refer to Godel, who is not the person that proved the theorem.

(III) Non sufficiency (of the associated description): in the Godel-example, "Godel" will not refer to Schmidt, who is the person that proved the theorem.

(IV) Incompatibility with the Kripkean picture: Kripke presents in N&N a very plausible picture of how the reference of a proper name is determined; this picture is incompatible with a descriptivist theory.

(V) Circularity: this objection applies to those DTs that appeal to meta-linguistic descriptions such as "the object named 'Socrates'".

2. A descriptivist theory

In addition to the distinction meaning-DT / reference-fixing-DT, there is another distinction among DTs that is lurking in the discussion in N&N: the description (or cluster) might be associated with the name for the entire community, or it might be so associated only for a specific user (or even: for a particular use by a specific user). We will call a DT of the first kind a *community-level* DT, and a theory of the second kind a *individual-level* DT.

We have then four kinds of DTs:

- (a) Community-level meaning DT
- (b) Community-level reference-fixing DT
- (c) Individual-level meaning DT
- (d) Individual-level reference-fixing DT

We will describe certain theory of type (d). As a reference-fixing individual-level theory of proper names our theory is a sort of descriptivist theory about proper names that holds that the (cluster of) descriptions associated with a name fix the name's reference but are not synonymous with

¹ Numbers within brackets refer to page numbers in Kripke(1980).

² See Wettstein(1989), p. 318.

it, and that each name has an associated cluster of descriptions for each individual speaker and each use of the name.

Our theory holds (*) true:

(*) All the semantic contribution a name does to what it is expressed by a sentence where

the name occurs is to provide certain individual.

Is (*) compatible with a descriptivist theory of proper names? A DT of the kind in (d) can be perfectly compatible with the direct reference view of proper names. An individual-use reference-fixing theory can claim that all the semantic function of a name is to provide an individual, while being at the same time sensitive to the intuition that in order to talk about an individual one needs to be able to intellectually reach the individual, or as we said above, to be able to have a cognitive fix on that individual. All a name does is to refer to some individual, but in order to be able to competently use a name a speaker must have a cognitive fix on the individual, and he can do so if the speaker in each one of his uses of the name has an associated 'description' (more in a moment about what exactly can be part of this 'description') that applies uniquely to the individual that is the referent of the name. Each person in a community might identify an individual in a different way, but if they all identify the same individual and are able to realize that they are all identifying the same individual they can agree to use certain expression as a name for that individual; even if each speaker is able to identify the referent of the name, whatever information he has associated with the name is not part of the meaning of the name, and need not be shared by other speakers. Being able to identify the referent of the name is just a condition on being able to understand the convention governing the use of the name (i.e. that people in the community have agreed to use such an expression as a name of such an individual), and this does not alter the fact that all the semantic function of the name is to provide some individual.

Now, there are some necessary conditions that are to be met if some person A is to be a competent user of a name *n* that refers to *x*:

One requirement, for instance, will be that A knows *n* as a syntactic object: that A is able to identify tokens of *n* under normal circumstances, that A knows *n*'s syntactic possibilities of combination, etc.

A must also (tacitly!) know that *n* is an expression whose meaning is of the kind that satisfies condition (*).

(In the spirit of what Evans calls 'Russell Principle' (see Evans(1982))): A must also know that the individual that *n* provides is *x*. In order to know this it is not necessary for A to have any special acquaintance with *x*. All that is required is that A is able to identify *x* some way or other. It might be enough if A has some description that uniquely picks up *x*, and so that A is able to 'intellectually reach' *x*, and so A is able to believe that *n* is a name of the object that is the so-and-so.

Different speakers might and most likely will, of course, use different 'descriptions' to identify the object they all refer to with the name *n*.

The information one speaker uses to identify the referent of a name will usually not be a single description, but a whole cluster of information –including not linguistically expressed information such as how a person looks like. Each such cluster of information that a subject binds together as identifying one single object is what we will call an *idea*. It might be that not all the 'descriptions' in an idea apply to the individual to which the idea applies. Also: not all the information in an idea has the same weight.

The idea associated with a name *n* will typically contain information such as 'the individual that those I learned the name *n* from referred to with *n*' or 'the individual that other people in this community refer to with *n*'.

Is not our proposal subject to the circularity objection we pointed out in (V) above? The kind of 'description' that one person *A* can not use to identify an individual *x* is the sort of description that appeals one way or the other to the fact that *A* is able to identify *x*. This would be circular. But it is not circular if the description appeals to what others are able to identify, and those are able to identify the object without appealing to what *A* is able to identify (in a direct or 'indirect' way: they would appeal indirectly to what *A* is able to identify if they were to appeal to someone else who in turn directly relies on what *A* is able to identify or directly relies on someone else who indirectly relies on what *A* is able to identify).

If an individual is already named α and I know it, then that the individual is named α will be part of my idea of that individual. It would not be possible for a group of us to first introduce some name β by identifying the individual that is to be called β as 'the individual that is called β ' (or 'that will be called β '). But, of course, if an individual is already named α , and I know that there is an individual named α , and if that is the only such individual, I can think of it as 'the only individual named α '. That this is possible and that there is no circularity involved here might be easier to see if we consider an analogous case where there is some convention involving some individual which is not a convention about the use of names. Suppose that a group of individuals have the convention of jumping twice every time they meet some particular individual. If I learn about this fact even without knowing anything else about the individual in front of whom people jump, I am able to intellectually reach this individual knowing that he is 'the one about whom there is the jumping-convention'. (Just with this information I would not, of course, be able to follow the jumping-convention because being able to follow this convention requires knowing who the individual is in a way that allows one to realize [in normal conditions] that one has met that particular individual. But in order to take part in a convention of naming something or someone what is needed is not to be able to realize whether one is in front of the object but simply to be able to intellectually reach

the object, and to do this it is enough that one has an idea that applies uniquely to that individual)³.

3. The Kripkean picture of proper names

The Kripkean picture of proper names is roughly the following: what a name refers to in one specific use is determined by a chain that goes from the object named to the speaker of that specific use; the chain starts with an initial 'baptism', where the name is introduced by some people (by naming the object either by ostension or by identifying it through some description); the chain then follows with those who learned it from the ones participating in the initial baptism, and those who learned from them, etc. What is essential for me to appropriately use the name "Feynman" is not that I appeal to some description such as "the one that introduced the theory of pair production and annihilation" that uniquely picks up Feynman independently of the community in which I am using the name; rather what it is essential is that I am appropriately placed in the chain that started with Feynman himself and an initial baptism and that has reached me through my community.

[One observation regarding what counts as 'my community': the community for the use of a particular name does not need to coincide with the community for most of the other words in my language (i.e. with a community such as the English speakers, the Portuguese speakers, etc). "Al Gore" is a name for Al Gore beyond the English speaking community. Is it the case that the community for a name is the whole of humanity? Certainly not, consider for instance the word "Monaco" which is a name in the Italian speaking community for the city that in German is called "München", whereas in, say, German "Monaco" is the name of a certain tiny European country. Also: A few of us might decide, for instance, to call Elisabeth Anscombe "Eliaams". Then my community for the name "Eliaams" will be just those of us who agreed to use "Eliaams" as a name for Anscombe.]

The erroneous picture that Kripke believes is behind a DT is the following⁴: someone alone in her room decides to use a certain name, e.g. "Gödel", to refer to, say, whoever proved that arithmetic is incomplete. Then, if the Schmidt-Gödel story is true it is clear that the person does not refer to Gödel when she tells herself, e.g. 'Gödel might have been a genius'. The name does refer to someone on that occasion, though: it refers to Schmidt.

Who does the person refer to when she uses the name when talking to other people (who took no part in her introduction of the name "Gödel" and who are in no other way connected with her in their use of "Gödel")? It is not clear what we should answer here. (Most of our puzzlement here arises, I believe, from the fact that the person would be acting in an irrational way,

³ For a more detailed discussion of the issues in sections 1 and 2 see Macià(1998).

⁴ He explicitly says so in Kripke(1980) p. 91.

since if she knows that “Gödel” is a name that she introduced while alone in her room, it would be irrational for her to try to use it in front of others that were not there in the room and who, she should believe, know nothing about this name). I would say that our intuitions are that if she refers at all she would refer to Schmidt not to Gödel. And whether we want to say that she refers or not it is clear that we must say that she does not communicate the thought she intends to express to the people who are listening to her. It seems clear that we must say that they do not understand what she wanted to say. Let’s look at another example that supports the same intuitions as this example, but which does not involve any consideration of descriptions.

Suppose there are two villages in a distant forest whose inhabitants speak the same language, say English, even though they are somewhat isolated from each other (say, by a mountain; the two villages, though, are only some 20 km apart). Suppose that a number of planes (say 20) start doing daily practical exercises in the aerial space around the mountain, and so can be seen by the people in the two villages. Assume that during the whole period of the practical exercises there is no communication at all between the people in one of the villages and the people in the other, and no communication at all between anyone in one of the villages and someone from somewhere else (in particular, people are not able to listen to the radio or use the internet). Suppose that in each of the villages people start recognizing the planes that take part in the exercises and give names to them. Suppose that, even if extremely unlikely to happen, it so happens that people in both villages give the same name to one of the planes, say they call it “Butterfly” (even though, as it should have been expected they give, of course, different names to the other 19 planes). Suppose now that one day two people A and B, one from each village, meet in the mountain and not realizing they are from different villages A tells B: “Butterfly did not do any flying exercises this morning”. Did B understand what A told her? I would say that she did not, even if she formed the same belief she would have formed had she really understood A. I think it is reasonable to require that for a speaker X to understand an utterance U by speaker Y it is necessary that X gains knowledge (by testimony) of what Y intended to say by uttering U (providing that X is justified in believing that Y is to be trusted as regarding the truth of what Y attempts to say by using U, and furthermore Y is to be trusted and what he intends to communicate is true). Let’s call this the *Transmission of Knowledge* requirement (TOK). TOK is not met in the case of the two villagers A and B meeting in the mountain: even if A’s utterance results in B forming a belief with the same content as the belief that A wanted to communicate, our intuitions are that B does not *know* that Butterfly had not been flying in the morning, even if B is completely justified in trusting A’s testimony, and furthermore A’s testimony is accurate. The reason that B does not acquire knowledge is that it is only by chance that she ends up forming the

right belief. Similarly, I do not gain the knowledge that it is now 5:23 pm by looking at a clock which, unknown to me, is broken and is always indicating '5:23 pm', even if now happens to be 5:23 pm.

Kripke's views might seem to contain the resources to explain why there is no transmission of information in the example of the two villagers A and B: they belong to two different and disconnected chains in the transmission of "Butterfly". Kripke could add: transmission of information using a proper name requires that people belong to the same chain for the name. Given that Kripke has this explanation for this example someone might even take this example to support Kripke's picture in front of a DT.

4. Coordination

As a challenge to the one that thinks that examples such as the one of the two villages supports Kripke's views in front of descriptivism we should point out that we can have examples of this sort when no proper name is involved, and when the expressions involved do clearly have a descriptivist content.

Consider the following case, similar to the one of the two villagers. There are two communities who speak different languages and who have never been in contact with each other. It so happens that by pure chance they use the same expression, say "kaumam", to express that it rains. Suppose that one day a man from one of the communities meets a woman from the other, and not knowing that they are from different communities tells her "Kaumam". As in the example in the previous section it seems clear that she did not understand him even if she formed the same belief she would have formed if she had understood what he said. And again the reason seems to be that the mechanism by which she formed the belief was not a reliable one. So in this case, again the TOK (transmission of knowledge) requirement is not met, and so we can say that the woman did not understand what the man said.

In this last example, though, we can not pretend to explain our intuitions appealing to a feature specific of proper names, since no proper name is involved. We could have used an example involving a very unlikely coincidence regarding ANY expression (for instance a sentence containing words such as "bachelor", "fortnight", or any other expression we wish).

I think that what this example shows is that, independently of the issue of what is the contribution that a proper name makes to what is expressed with the use of the proper name, or even independently of what makes it the case that a proper name makes the contribution that it makes to what it is expressed with a sentence containing it, there is one general requirement that proper names *as well as any other sort of expression* must meet in order for speakers to be able to communicate using those expressions.

The requirement is that any two speakers that can successfully communicate using an expression α must be, as we will say, *coordinated* in

their use of α . Roughly: the notion of 'being coordinated in the use of an expression' attempts to capture the idea of using an expression in a way that, unlike what was the case in the examples we considered above, is not independent or disconnected from what others do. First we will give a characterization of this notion for the particular case where the expression under consideration is a proper name, and then we will suggest how this characterization could be generalized.

I am *coordinated* with an individual A in my possession of the name α to refer to x if and only if I am immediately coordinated with A or I am coordinated with someone who is immediately coordinated with A. I am *immediately coordinated* with A in my possession of the name α to refer to x if one of the following obtains: (1) I learned the name (at least in part) from A. That is, I learned from A that the expression α is used in some community as a name in accordance with (*) to determine x. The qualification "at least in part" is meant to include also the case where I realized that there is the relevant kind of convention by observing how some individuals that included A were using the expression α , and to include also the case where by observing A I would have *some* evidence that there is the relevant kind of convention but not enough evidence as to allow me to *know* that α is used to determine x in accordance to (*), and so I also rely in other sources of evidence --each of which by itself not being enough as to provide me with knowledge of the relevant convention, even though taken all together they do provide me with enough evidence as to give me *knowledge* of the convention; (2) A learned the name (at least in part) from me, (3) We both took part in the introduction of the name. We might have done so by explicitly saying "let's use α as a name of such-and-such individual". That such-and-such individual might have been specified by means of some descriptions, or might have been identified using a demonstrative and relying on the fact that each of us were able to identify x and that furthermore each of us knew that the other was identifying the same individual as he did. Or we can both be part of the introduction of a name if we agreed on following certain general rule for the introduction of names (like the one governing the assignment of number-names to the streets in certain cities). There could possibly be still other ways of two people participating together in the introduction of a name.

We can try to give a rough generalization of this characterization to any sort of (simple) expression: we could say that I am *coordinated* with an individual A in my possession of an expression β to mean y if and only if I am immediately coordinated with A or I am coordinated with someone who is immediately coordinated with A. I am *immediately coordinated* with A in my possession of the expression β to mean y if one of the following obtains: (1) I learned the expression (at least in part) from A. That is, I learned from A that the expression β is used in a certain community in some specific way to contribute certain sort of meaning to the complex expressions where it appears, and that

furthermore β has such-and-such syntactic category and enjoys such-and-such possibilities of combination with other expressions. (2) A learned the expression β (at least in part) from me, (3) We both took part in the introduction of the expression. We might have done so by explicitly saying "let's use β as an expression of such-and-such syntactic category to make such-and-such contribution to what it is expressed by the more complex expressions that contain it. There could possibly be still other ways of two people participating together in the introduction of an expression.

In the light of the previous discussion it is clear that what Kripke regarded as the erroneous picture about proper names is not just erroneous about proper names but about any sort of expression at all. No one can communicate using an expression that she has endowed with meaning by an act performed when alone in one's room.

That the general picture about naming that Kripke presents is correct is what it should be expected if the descriptivist theory that we defend is true. That this is so follows from (1) the fact that an idea (our equivalent for a 'cluster of descriptions') can contain information such as 'the one that people from whom I learned that α is a name referred to by α ', 'the one that other people in such a community call α ', 'the one that such and such people call α ', etc. (2) The fact that to people need to be coordinated in order to understand each other when using a name.

From (1) it follows that if I successfully use a name to communicate then it must be the case that there is a chain that links me to the introduction of the name. This is so also for any other word, though. And so it is not something that distinguishes proper names from other expressions. In particular it is not something that would distinguish proper names from expressions that work like descriptions do.

Fact (2) accounts for what does distinguish proper names from other expressions. Notice that I would not count as knowing what "fortnight" means if all I knew was that it means 'what those from whom I learned that "fortnight" is a common noun meant when they used "fortnight"'. Still it seems that it is possible to have the equivalent of just that information about a proper name and be able to use it successfully to communicate. The difference comes from the fact that proper names are special in that they meet (*). So for the case of a proper name α the general necessary condition that I be coordinated in my use of the name, and so that I must have learned from others that α is name, already makes it possible to me to know all it is necessary in order to be able to successfully use the name.

Notice that the fact that the general picture about naming that Kripke presents is correct is what it should be expected if the descriptivist theory that we defend is true, makes it the case that the intuitive plausibility of the

kripkean picture can not be used at all as an argument against our specific kind of DT⁵.

REFERENCES

- Evans, Gareth: "The Causal Theory of Names", *Aristotelian Society Supplementary Volume*, 47, pp. 187-208, (1973). Reprinted in Evans, G.: *Collected Papers*, pp.1-24. Oxford, Clarendon Press. (1985)
- Evans, Gareth: *The Varieties of Reference*, Oxford, Clarendon Press. (1982)
- Kripke, Saul: *Naming and Necessity*. Cambridge, Mass., Harvard University Press. (1980)
- Macià, Josep: "Does Naming and Necessity Refute Descriptivism?", *Theoria*, 13-3, pp. 445-476. (1998)
- Stalnaker, Robert: "Assertion", In Peter Cole (ed.), *Syntax and Semantics*, vol 9: *Pragmatics*, pp. 315-332. New York: Academic Press. (1978)
- Wettstein, H: "Turning the Tables on Frege or How is it that "Hesperus is Hesperus" is trivial?", *Philosophical Perspectives*, 3. (1989)

Josep Macià
 University of Barcelona
 e-mail: josep@alum.mit.edu

⁵ In writing this paper I have benefited from discussions with many people, including Stephen Neale, Manuel Pérez, Joseph Almog, Ignasi Jané, Ignacio Vicario, Robert Stalnaker and Manuel García-Carpintero. I wish to thank them. Research for the paper has been funded by grants from the Catalan Government's CIRIT (SGR97-396), and the Spanish Government's DGICYT (PB96-1091-C03-03).

EL ANÁLISIS RUSSELLIANO DE LAS EXPRESIONES DENOTATIVAS Y LA TEORÍA MEDIEVAL DEL DESCENSO

PALOMA PÉREZ-ILZARBE
Universidad de Navarra

ABSTRACT: The horizon of this paper is the question "what are our assertions *about*?". More specifically, I examine two famous theories on the subject: the theory of denoting phrases found in Russell's *The Principles of Mathematics*, and the medieval theory of *descensus*. The first section contains an account of the Russellian theory, and the second section is devoted to the medieval theory, illustrating the close similarity between the two. The third and final section presents two shortcomings that make both theories in some sense incomplete: they seem incapable of dealing with unitary and null classes.

KEYWORDS: Denoting, *descensus*, *suppositio*, quantification, unitary class, null-class.

Entre los objetivos de este congreso se proponía aprovechar el cambio de milenio "como una ocasión simbólica para constatar el estado actual y las perspectivas de futuro de la Filosofía Analítica". En mi opinión, sería una buena idea completar ese examen del presente y el futuro de la filosofía analítica volviendo también la mirada hacia el pasado. Pienso que sí puede decirse, con el permiso de Acero, que la filosofía analítica no es "agua pasada"¹, es porque ha sabido recuperar la perspectiva histórica y el trabajo cooperativo como medio para alcanzar una mejor comprensión de los problemas filosóficos². Compartiendo este espíritu, voy a poner en relación dos teorías muy distantes en el tiempo (muy distantes entre sí, y bastante distantes ya del momento presente), con la esperanza de que su examen ayude a iluminar un problema que nunca dejará de interesar a la filosofía del lenguaje:

¹ Ver la opinión contraria en J. J. Acero, *Filosofía y análisis del lenguaje*, Cincel, Madrid, 1989.

² Ver J. Nubiola, *La renovación pragmatista de la filosofía analítica. Una introducción a la filosofía contemporánea del lenguaje*, Eunsa, Pamplona, 1994.

¿de qué hablamos con nuestros enunciados?

Para investigar posibles respuestas a la pregunta, tomaré, por una parte, a uno de los pioneros en el análisis filosófico, Bertrand Russell (1872-1970), y atenderé a una teoría temprana y pronto abandonada, la teoría del denotar tal como es expuesta en el capítulo V de *Los principios de la matemática* (1903, 1938). Por otra parte, retrocederé algunos siglos más hasta la filosofía del lenguaje medieval, y consideraré en concreto la tradición terminista (que nace en el siglo XII, florece en el XIV y continúa viva hasta el XVI)³, llamada así porque se ocupa del análisis de las propiedades de los términos⁴. La doctrina que examinaré dentro de esta tradición es la llamada teoría del descenso (*descensus*)⁵, y trataré de establecer un paralelismo entre el análisis de los enunciados cuantificados operado mediante el descenso y el ofrecido por Russell en *Los principios de la matemática*⁶. Terminaré denunciando los límites de ambas teorías, a través de dos casos concretos con los que ninguna de ellas logra manejarse.

1. Las expresiones denotativas en *Los principios de la matemática*

En el prefacio de esta obra, Russell explica cómo un problema inicial relativo a la filosofía de la dinámica le condujo hasta la lógica simbólica, en un intento de descubrir el significado de la palabra "cualquier" ("*any*") (*PoM*, xvi)⁷. Del significado de esta palabra, junto con el de "todos" ("*all*"), "cada"

³ Por supuesto, ni la filosofía medieval del lenguaje en general ni la lógica terminista en particular conforman una tradición única y homogénea, sino que en su seno se desarrollan muchas escuelas y tendencias que luchan entre sí, se entrelazan, se corrigen, se completan... Pero la tradición escolástica puede verse, en mi opinión, como una tradición porque las distintas escuelas comparten un mismo carácter (el de la lógica como ciencia del lenguaje), origen (los problemas surgidos en la interpretación de textos teológicos) y propósito (el análisis científico del lenguaje y la explicitación de sus presupuestos). Sobre el sentido de la lógica medieval se puede consultar, por ejemplo, E. A. Moody, "The medieval contribution to logic", *Studies in medieval philosophy, science and logic*, University of California Press, Berkeley, 1975, 371-392. Sobre el origen de la lógica terminista, L. M. de Rijk, *Logica modernorum. A contribution to the early terminist logic*, Van Gorcum, Assen, 1962, 1967.

⁴ Aunque la etiqueta "filosofía analítica" sea un invento reciente, también puede aplicarse a esta tradición, que en mi opinión constituye una muestra extraordinaria de hasta dónde se puede llegar en el análisis minucioso del lenguaje.

⁵ No estudiaré la teoría de ningún autor en concreto, sino que expondré los rasgos básicos de la teoría estándar del descenso.

⁶ La comparación entre la teoría russelliana del denotar y la interpretación medieval de los términos generales cuantificados ya ha sido hecha por Geach, aunque él no alude explícitamente a la doctrina del descenso, sino sólo a la teoría de la suposición. Ver P. T. Geach, *Reference and generality. An examination of some medieval and modern theories*, Cornell University Press, Ithaca, 1980, en especial las páginas 73-104.

⁷ Ver R. Hursthouse, "Denoting in 'The principles of mathematics'", *Synthese* 45 (1980), 40: el problema que preocupa a Russell es el de cómo algo esencialmente predicativo puede

("every"), "un" ("a") y "algún" ("some")⁸ se ocupa en el capítulo V, el conocido capítulo sobre el denotar⁹. Russell llama a las expresiones que contienen estas palabras "expresiones denotativas" ("*denoting phrases*"), aunque ellas mismas no denotan, sino que *indican* o *expresan* conceptos denotativos. Son éstos los que están en una relación especial con ciertos términos, la relación lógica de *denotar*¹⁰.

Antes de explicar en qué consiste esta relación, Russell destaca su interés, señalando que la relación de denotar es la que hace posible la *descripción*: gracias al denotar podemos, mediante el uso de conceptos, designar una cosa que no es un concepto (*PoM*, § 56). Al recordar la división russelliana de los términos en cosas y conceptos, se comprende el alcance de esta observación: los conceptos, que normalmente se usan como predicados para hacer una aserción acerca de un sujeto, pueden también ser usados como sujeto, tomando una variedad de formas gramaticales ("hombre", "un hombre", "algún hombre", "cualquier hombre", "cada hombre", "todos los hombres" y sus equivalentes). Pero la naturaleza de las cosas y la de los conceptos es distinta, y una de las marcas de esta distinción es la capacidad de denotar, que corresponde exclusivamente a los conceptos (*PoM*, § 48).

El problema que plantea el uso de conceptos como sujeto es, para Russell, el del "acerca de". Una proposición como la expresada por "Sócrates es un hombre" es claramente acerca de Sócrates (la cosa que es uno de los constituyentes de la proposición), pero una proposición como la expresada por "Encontré un hombre" no es acerca de *un hombre* (el concepto que es uno de los constituyentes de la proposición). En *Los principios de la matemática*, Russell cree que la relación lógica de denotar resuelve el problema del "acerca de".

¿Por qué interesa a Russell el "acerca de"? Según Peter Hylton, una de las preocupaciones de Russell cuando introduce los conceptos denotativos es

ocupar el lugar de un término en una proposición.

⁸ Dejo de lado, como hace inicialmente Russell, el estudio de la partícula "*the*".

⁹ Para quien prefiera un resumen del texto de Russell, puede consultar C. E. Cassin, "Russell's discussion of meaning and denotation: a re-examination", *Essays on Bertrand Russell* (E. D. Klemke, ed.), University of Illinois Press, Urbana, 1970, 256-272.

¹⁰ No debe confundirse el "término" russelliano con el sentido medieval de "término" ("*terminus*": signo ponible en una proposición). Como se sabe, para Russell los términos no son palabras sino las entidades de su ontología: "cualquier cosa que puede ser un objeto de pensamiento, o que puede ocurrir en cualquier proposición verdadera o falsa, o que puede contarse como *uno*" (*PoM*, § 43). Las proposiciones son también ciertas entidades especiales en la ontología russelliana, no lingüísticas, a diferencia de las proposiciones en el sentido medieval (ver nota 19). Ver P. Hylton, *Russell, idealism and the emergence of analytic philosophy*, Clarendon Press, Oxford, 1990, 171-172.

la cuestión de cómo podemos *entender* las proposiciones¹¹. Si entender una proposición requiere un contacto directo (*acquaintance*) de la mente con las cosas de las que habla la proposición, surgiría el problema de cómo podemos entender proposiciones que tratan acerca de un número infinito de objetos. El concepto denotativo aparece como un intermedio entre la expresión lingüística y ese conjunto infinito de objetos, intermedio del que sí puede tenerse un conocimiento directo¹².

Russell no define la relación de denotar. Simplemente apunta que se trata de una relación lógica que se da entre algunos conceptos y ciertos términos, tales que cuando al figurar el concepto en una proposición ésta no es *acerca de* ese concepto, sino acerca del término (*PoM*, § 56). Más adelante ya no habla de "un término", sino de "un objeto", que es una especial combinación de términos (*PoM*, § 58). Russell examina las distintas expresiones denotativas para tratar de descubrir si hay una única manera de denotar distintos tipos de objetos, o si se trata, en cambio, de distintos modos de denotar¹³.

Seguiré el análisis de Russell de los cinco "cuantificadores": "todos", "cada", "cualquier", "un", "algún". Estas expresiones, al añadirse a un concepto-clase, componen una expresión denotativa que, según el Russell de *Los principios de la matemática*, siempre denota (en rigor, como he dicho, lo que siempre denota es el concepto denotativo expresado por ella). Para ilustrar la diferencia entre los objetos denotados por cada uno de los conceptos denotativos, Russell propone tomar un concepto-clase que defina una clase con un número finito de términos. En este caso, los objetos denotados pueden indicarse mediante listas conjuntivas o disyuntivas:

- (1) *Todos los a* denota a_1 y a_2 y ... y a_n .
- (2) *Cada a* denota a_1 y denota a_2 y ... y denota a_n .
- (3) *Cualquier a* denota a_1 o a_2 o ... o a_n , donde "o" significa que es irrelevante cuál tomemos.
- (4) *Un a* denota a_1 o a_2 o ... o a_n , donde "o" significa que no debe tomarse

¹¹ P. Hylton, "Beginning with analysis", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 197.

¹² Ver también H. Noonan, "The 'Gray's Elegy' argument —and others", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 77.

¹³ Nicholas Griffin explica cómo esta preocupación de Russell en *Los principios de la matemática* se comprende mejor cuando se sabe que en un escrito inédito de 1901 Russell distinguía cinco maneras distintas de denotar. En *Los principios de la matemática* ha decidido que se trata de una única relación con objetos distintos (ahora seis, puesto que incluye "the"). N. Griffin, "Denoting concepts in *The principles of mathematics*", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 29.

ninguno en particular.

- (5) *Algún* a denota a_1 o denota a_2 o ... o denota a_m , donde no es irrelevante cuál se tome, sino que debe tomarse un a en concreto.

En cada caso el concepto denotativo denota un objeto distinto. Estos objetos son de una naturaleza peculiar: se trata de distintas combinaciones de términos. Excepto en el primer caso, en que lo denotado son muchos términos, en los restantes casos no es ni uno ni muchos, sino una especial combinación¹⁴. Russell inventa para ellas los siguientes nombres¹⁵:

a) Lo denotado por *todos los* a es una **conjunción numérica**: la colección de todos los a tomados colectivamente, todos a la vez. Decir que "Todas las monedas de mi bolsillo suman 200 pesetas" sería algo así como decir que esta y esta y esta moneda, en conjunto, suman 200 pesetas.

b) Lo denotado por *cada* a es una **conjunción proposicional**: la colección de los a , pero en este caso tomados distributivamente, uno a uno. Decir que "Cada una de las monedas de mi bolsillo es de 25 pesetas" es decir que esta moneda es de 25 y esta otra es de 25 y esta otra...

c) Lo denotado por *cualquier* a es una **conjunción variable**¹⁶: según Russell, el concepto denota "un a variable". Esto quiere decir que la proposición en la que aparece no habla de ningún a en particular, aunque será verdadera de uno cualquiera que se tome (y sin embargo, no será verdadera de todos). "Puedes coger cualquier moneda de mi bolsillo": no puedes cogerlas todas, pero tampoco hay una en particular que sea la que puedes coger, sino que para cualquiera que tomes es verdadero "puedes coger ésta".

d) Lo denotado por *un* a es una **disyunción variable**: tampoco la proposición habla de un a en particular y en este caso ni siquiera es verdadera de ninguno cualquiera que se tome, aunque es verdadera de uno u otro. "Debes coger una moneda de mi bolsillo": no es esta, ni esta, ni esta moneda concreta la que debes coger, pero es cierto que debes coger una u otra de ellas.

e) Lo denotado por *algún* a es una **disyunción constante**: ahora sí que hay un cierto compromiso con un a en particular, aunque ese particular puede ser cualquiera. "Alguna moneda de mi bolsillo es de 25 pesetas": de las

¹⁴ Una combinación, indica Russell, que se realiza sin usar relaciones (*PoM*, § 59).

¹⁵ Para refrescar la memoria de quienes ya conozcan el texto de Russell y facilitar la comprensión a quienes no lo conozcan, no usaré los ejemplos russellianos, sino otros de mi invención en los que creo que queda clara la distinción entre las diferentes combinaciones de objetos.

¹⁶ Russell usa el nombre de "conjunción" aunque caracteriza el objeto denotado mediante una disyunción. Lo hace así porque considera que el tipo de combinación es distinto al de cualquiera de las formas de disyunción, y le parece más bien algo intermedio entre una conjunción y una disyunción (*PoM*, § 59).

monedas de mi bolsillo, ciertas monedas determinadas son de 25 pesetas (no cualquier moneda que tomemos lo será).

Es importante recordar que estos objetos son combinaciones distintas de los mismos términos. Para cada concepto-clase, el análisis russelliano del "acerca de" termina en las mismas cosas, aunque combinadas de distinta manera. Al decir "Encontré un hombre", asegura Russell, "toda la raza humana está contenida en mi aserción" (*PoM*, § 62). Las distintas combinaciones de términos en que consisten los distintos objetos denotados vendrían a ser distintas maneras de involucrar a la raza humana en las aserciones. Sin recurrir a denotaciones, ni a objetos, ni a combinaciones de términos, la teoría medieval del descenso también explica cómo, si decimos "Todo hombre corre" o si decimos "Algún hombre corre", hablamos de los mismos hombres, pero de distinta manera.

2. La doctrina del *descensus* en la tradición terminista

El análisis de las propiedades de los términos constituye la aportación original al estudio del lenguaje de esta tradición medieval, también llamada *logica modernorum* (la lógica de los "modernos", en contraste con la lógica "antigua", que consistía básicamente en comentarios a las obras de Aristóteles y Porfirio¹⁷). Como resultado de su minucioso análisis de los enunciados, los lógicos terministas elaboran una teoría del significado y de la referencia, centrada en la propiedad de los términos conocida como suposición (*suppositio*).

Aunque es difícil caracterizar la suposición en pocas palabras y de forma unívoca, ya que la noción tiene una compleja historia¹⁸, trataré de hacer una caracterización general para quienes no estén familiarizados con ella. La suposición es la propiedad que tienen los términos de *estar por* las cosas en las proposiciones¹⁹. La idea del *estar por* tiene su origen en el descubrimiento de

¹⁷ Ver J. Pinborg, *Logica e semantica nel medioevo*, Boringhieri, Torino, 1984, 21-22.

¹⁸ Ver L. M. de Rijk, "The origins of the theory of the properties of terms", *The Cambridge history of later medieval philosophy* (N. Kretzmann et al, eds.), University Press, Cambridge, 1982, 161-173; D. P. Henry, "The early history of suppositio", *Franciscan studies* 23 (1963), 205-212; O. Ducrot, "La théorie médiévale de la supposition", *History of linguistic thought and contemporary linguistics* (H. Parret, ed.), de Gruyter, Berlin, 1976, 189-277; E. J. Ashworth, "The doctrine of supposition in the sixteenth and seventeenth centuries", *Archiv für Geschichte der Philosophie* 51 (1969), 260-285. Un resumen de la evolución de la noción se encuentra en mi *El significado de las proposiciones. Jerónimo Pardo y las teorías medievales de la proposición*, Eunsa, Pamplona, 1999, 43-50.

¹⁹ A partir de aquí, "proposición" se usará en el sentido técnico que tiene en la tradición escolástica, como un tipo especial de oración compleja. La caracterización de este tipo especial de signo lingüístico puede hacerse tomando como base las nociones aristotélicas de *protasis*

una doble dimensión en la significación de los nombres: intensional (significación de una naturaleza universal) y extensional (significación de las cosas concretas que participan de esa naturaleza universal). Un nombre puede, por el hecho de ser significativo, traer a la mente un cierto significado, pero cuando se usa en una proposición se toma ese nombre por unos individuos concretos. Por ejemplo, en "Algún hombre corre", el término "hombre" *está por* ciertos hombres concretos.

El contexto proposicional determina distintos tipos de suposición, de los que no puedo ocuparme aquí²⁰. Me interesan ahora los distintos modos de suposición personal, que es el tipo de suposición que tienen los términos comunes cuando se toman por los individuos que caen bajo ellos: por ejemplo, la suposición que tiene el término "hombre" en "Todo hombre corre", en "Algún hombre corre" o en "Todo el que corre es hombre". Los distintos modos de suposición personal nos dan las distintas relaciones que puede tener el término universal respecto a los individuos particulares.

Lo que está en juego aquí, como en la teoría russelliana del denotar, es el problema del "acerca de". Aunque en una proposición aparezcan términos universales, lo que la proposición dice no es acerca del universal en abstracto, sino acerca de los individuos concretos que caen bajo él. La herramienta construida por los lógicos escolásticos para poner de manifiesto los distintos modos en que las proposiciones con términos generales hablan de los correspondientes individuos es la doctrina del descenso²¹. El descenso es una inferencia desde la proposición en la que aparece el término general hasta proposiciones en las que aparecen términos singulares, conectados mediante conjunciones y disyunciones. La sustitución del término general por los singulares, según los distintos modos de conexión, es el modo de hacer explícita la relación entre el término universal que aparece en una determinada

(enunciado afirmativo o negativo de algo acerca de algo), o de *logos apophantikos* (oración en la que se da la verdad y la falsedad).

²⁰ Para tener noticia de ellos, puede consultarse P. Boehner, *Medieval logic. An outline of its development from 1250 to c. 1400*, University Press, Manchester, 1966, 27-51. Allí Boehner compara explícitamente la teoría medieval de la suposición con la teoría moderna de la cuantificación.

²¹ Se puede encontrar una presentación general de la teoría medieval del descenso en P. Boehner, *Medieval logic*, 29-31 y 41-43; y de sus prolongaciones postmedievales en E. J. Ashworth, *Language and logic in the post-medieval period*, Reidel, Dordrecht, 1974, 207-223. Ver también mi artículo "The doctrine of descent in Jerónimo Pardo: meaning, inference, truth", *Studies on the history of logic. Proceedings of the III. Symposium on the History of Logic* (I. Angelelli y M. Cerezo, eds.), de Gruyter, Berlin, 1996, 173-186, y *El significado de las proposiciones*, 50-56.

proposición y los individuos de los que ésta habla²².

Cuáles son los individuos de los que habla la proposición y cómo lo hace depende de la suposición con que está tomado el término general, y esto a su vez depende del contexto proposicional (la manera en la que los "cuantificadores" afectan a los términos determina su modo de suposición personal). Los cuatro tipos de suposición personal quedan definidos por otros tantos tipos de descenso, que explicita los individuos particulares de los que habla cada proposición, indicando los distintos modos de hablar de ellos:

a) Cuando un término está afectado por un cuantificador particular, si no hay otro signo que lo modifique, el término se toma en suposición **determinada**, como "hombre" en "Algún hombre es blanco". Para explicitar los individuos de los que habla la proposición, se lleva a cabo el descenso: "Algún hombre es blanco, por tanto, este hombre₁ es blanco o este hombre₂ es blanco o ... o este hombre_n es blanco". Este tipo de descenso recibe el nombre de descenso **disyuntivo** (se desciende desde la proposición con el término general a una disyunción de proposiciones con términos singulares).

b) Cuando un término está afectado mediatamente por un cuantificador universal, como "animal" en "Todo hombre es animal", el término se toma con suposición **meramente confusa**. Para determinar los individuos de los que habla la proposición habrá que efectuar un descenso **disyunto**: "Todo hombre es animal, por tanto, todo hombre es o bien este animal₁ o bien este animal₂ o ... o bien este animal_n" (se desciende desde la proposición con el término general hasta una proposición cuyo predicado es una disyunción de términos singulares).

c) Cuando un término está afectado inmediatamente por un cuantificador universal, como "hombre" en "Todo hombre es animal", el término se toma con suposición **confusa y distributiva**. El descenso que debe efectuarse para explicitar los individuos de los que habla la proposición se denomina descenso **copulativo**: "Todo hombre es animal, por tanto, este hombre₁ es animal y este hombre₂ es animal y ... y este hombre_n es animal" (se desciende a una conjunción de proposiciones con términos singulares).

d) Cuando un término está afectado inmediatamente por un cuantificador universal tomado colectivamente, como "apóstoles" en "Todos los apóstoles son doce", el término tiene un tipo especial de suposición, que podemos llamar **colectiva**. El descenso correspondiente toma el nombre de descenso **copulado**:

²² Para una presentación del análisis terminista de las proposiciones como una teoría de la referencia proposicional, ver N. Kretzmann, "Medieval logicians on the meaning of the propositio", *The journal of philosophy* 67 (1970), 768-771.

"Todos los apóstoles son doce, por tanto, este apóstol₁ y este apóstol₂ y ... y este apóstol₁₂ son doce" (se desciende a una proposición cuyo sujeto es una conjunción de términos singulares).

El resultado de cada uno de estos cuatro tipos de descenso puede ponerse en correlación con alguno de los objetos denotados por los conceptos denotativos de Russell:

a) Mediante el descenso disyuntivo se llega a algo muy parecido a la disyunción constante russelliana: la proposición "Algún hombre es blanco" habla de los hombres de un modo especial, como afirmando implícitamente la verdad de alguna de las proposiciones singulares de la forma "Este hombre_n es blanco".

b) Mediante el descenso disyunto se llegaría a una disyunción variable: la proposición "Todo hombre es animal" no se compromete con la verdad de ninguna proposición particular de la forma "Todo hombre es este animal_n", pero sí afirma que todo hombre es un animal u otro.

c) Mediante el descenso copulativo se llega a una conjunción proposicional: la proposición "Todo hombre es animal" se comprometería con la verdad de todas y cada una de las proposiciones singulares "Este hombre_n es animal".

d) Mediante el descenso copulado se llega a lo que Russell llama conjunción numérica: quien afirma la proposición "Todos los apóstoles son doce" no se compromete con la verdad de ninguna proposición singular de la forma "Este apóstol_n es doce", sino que de lo que afirma que son doce es de los doce apóstoles tomados en conjunto²³.

Este paralelismo entre relación de denotar y descenso podría quedarse en una simple curiosidad, de no ser por la existencia de algunos problemas lógicos con los que ambas teorías se encuentran, y que sirven para marcar los límites de este tipo de análisis de las proposiciones. En la sección final voy a ocuparme de dos de ellos²⁴.

3. Los problemas

En el capítulo VI de *Los principios de la matemática*, a propósito de la

²³ La conjunción variable de Russell (ese "intermedio entre una conjunción y una disyunción": ver nota 16) es la que se queda sin correlato en el cuadro de los tipos de descenso, puesto que las posibilidades son estas cuatro: disyunción de proposiciones, disyunción de términos, conjunción de proposiciones, conjunción de términos.

²⁴ Geach, por ejemplo, atiende a otros problemas, haciendo una crítica de estas teorías desde la interpretación de los cuantificadores en la tradición fregeana. Ver P. T. Geach, *Reference and generality*, 105-135.

noción de *clase*, Russell debe hacer frente a ciertas dificultades de interpretación suscitadas por unos conceptos denotativos peculiares: los que proceden de conceptos-clase que denotan un único término, y los que proceden de conceptos-clase vacíos. La doctrina del descenso se encuentra con dificultades semejantes.

La existencia de clases unitarias obliga a Russell a afirmar que la noción de *y* no forma parte del significado de una clase (*PoM*, § 71), aunque un poco antes haya definido la clase como una conjunción numérica de términos (*PoM*, § 67). Cuando una clase contiene sólo un término, la clase se identifica con ese único término *y*, por tanto, lo que denota el concepto denotativo correspondiente es el término mismo, y no una especial combinación de términos. Esto parece entrar en conflicto con la convicción russelliana de que no hay distintos modos de denotar, sino distintos tipos de conceptos denotativos y de objetos denotados: para los conceptos-clase que denotan un único término, todos los conceptos denotativos denotarán el mismo objeto, de modo que se diluye la supuesta distinción entre ellos.

Algo parecido sucede en la teoría del descenso, que está diseñada para tratar con términos generales que tienen una pluralidad de singulares bajo ellos, pero que no está tan preparada cuando el término general tiene un único singular. ¿Qué tipo de descenso podría efectuarse bajo la proposición "Todo dios es omnipotente", si sólo hay una proposición singular a la que descender? Y, si el modo de suposición viene dado por el tipo de descenso, ¿cómo distinguir la suposición que tiene el término "dios" en "Todo dios es omnipotente" de la que tiene en "Algún dios es omnipotente"?²⁵

También es un problema para estas dos teorías el tratar con lo que Russell denomina conceptos-clase vacíos. Para Russell es indudable que no hay tal cosa como la clase vacía (puesto que una clase queda definida por sus términos: si no hay términos, no hay clase), pero sí hay conceptos-clase vacíos y conceptos de una clase vacíos²⁶. Aunque para el Russell de *Los principios de la matemática* un concepto denotativo siempre denota (*PoM*, § 58), hay conceptos que denotan aunque no denotan nada (*PoM*, § 73).

Russell propone la siguiente definición de un concepto que denota, pero no denota nada: los conceptos denotativos asociados con un concepto-clase *a* no denotan nada cuando y sólo cuando "*x* es un *a*" es falso para todos los valores de *x*. Análogamente, en la lógica terminista hay términos que no

²⁵ Matthews se hace estas preguntas a propósito de la teoría del descenso que Ockham expone en su *Summa logicae*, y declara que no puede encontrar en el texto una respuesta: G. B. Matthews, "Suppositio and quantification in Ockham", *Noûs* 7 (1973), 21.

²⁶ Para la distinción entre concepto-clase y concepto de una clase, ver *PoM*, § 67.

suponen, y se reconocen cuando no hay ningún objeto del que sea verdadero decir "este es un *a*". Por ejemplo, en "Algún hombre que es caballo corre", el término "hombre que es caballo" no supone por nada, porque no hay nada de lo que sea verdadero decir "este es un hombre que es caballo".

El problema lógico que suscitan los conceptos denotativos que no denotan es el de cómo interpretar las proposiciones que los contienen. "Las quimeras son animales" parece una proposición verdadera, y parece que el concepto denotativo *las quimeras* denota, puesto que la proposición no es acerca de él mismo. Sin embargo, el concepto no denota nada, porque la función proposicional "*x* es una quimera" es falsa para todos los valores de *x*. Pero si no hay quimeras cuya combinación pudiera ser denotada por el concepto, ¿acerca de qué es, entonces, la proposición? Russell ensaya distintas estrategias para interpretar la proposición (mediante una implicación formal: "*x* es una quimera implica *x* es un animal para todos los valores de *x*"; intensionalmente, como relación de predicados: el adjetivo *animal* es parte de la definición del adjetivo *quimérico*; extensionalmente, como relación entre denotaciones: nada es denotado por *una quimera* que no sea denotado por *un animal*), pero ninguna le satisface: todas estas proposiciones *implican* que las quimeras son animales, pero no se trata de *la misma* proposición. Russell concluye diciendo que lo más sensato sería rechazar por completo la proposición que causa tales problemas (*PoM*, § 73). La teoría de la denotación, que establece una semántica independiente para las frases cuantificacionales²⁷, se encuentra, en el caso de los conceptos-clase vacíos, con expresiones a las que esa semántica no puede aplicarse.

Parejamente, las quimeras también fueron motivo de preocupación para los lógicos medievales²⁸, y fuente de problemas para la teoría del descenso²⁹. Cuando un universal no tiene individuos singulares bajo él, no cabe efectuar ningún descenso, pues ¿a qué individuos señalarían los demostrativos en las proposiciones singulares "Esta quimera₁ es animal y esta quimera₂ es animal y ... y esta quimera_n es animal"? Y ocurre lo mismo para cualquier caso en que el

²⁷ La alternativa es no buscar una semántica propia, sino disolver esas frases mediante cuantificadores, como es costumbre en la tradición fregeana, incluyendo al mismo Russell en escritos posteriores.

²⁸ La accidentada vida de la quimera desde Homero hasta Buridán es descrita por S. Ebbesen en "The chimera's diary", *The logic of being* (S. Knuuttila and J. Hintikka, eds.), Reidel, Dordrecht, 1986, 115-143. Sobre el tratamiento de las quimeras en la lógica escolástica postmedieval, ver E. J. Ashworth, "Chimeras and imaginary objects: a study in the post-medieval theory of signification", *Vivarium* 15 (1977), 57-79.

²⁹ Ver E. Karger, "Consequences et inconsequences de la supposition vide dans la logique d'Ockham", *Vivarium* 16 (1978), 46-55.

término general no suponga, aunque esa falta de suposición sea accidental (por ejemplo, "hombre" en "Algún hombre corre", en el supuesto de que no hubiera hombres). Hay algunas proposiciones que para la teoría del descenso, como para la teoría del denotar russelliana, resultan imposibles de interpretar.

Termino recordando la frase citada de Russell "toda la raza humana está involucrada en mi aserción". Lo que parece proponer tanto la teoría russelliana como la medieval es algo así como una *cuantificación restringida* a un cierto dominio. Es el sujeto el que indica el dominio de objetos del que deben tomarse los valores de las variables a las que se aplica el predicado, y de ahí la exigencia de presuponerlo no vacío³⁰. Pero de hecho hay conceptos-clase vacíos y de hecho hay términos que no suponen, y ninguna de las dos teorías alcanza a dar cuenta de las proposiciones en que aparecen³¹. El problema del "acerca de" sigue abierto, y eso quiere decir que, en el siglo XXI, quienes nos dedicamos al análisis del lenguaje tendremos un interesante campo en el que investigar.

BIBLIOGRAFÍA

- Acero, Juan José: *Filosofía y análisis del lenguaje*, Cincel, Madrid, 1989.
- Ashworth, E. J.: "The doctrine of supposition in the sixteenth and seventeenth centuries", *Archiv für Geschichte der Philosophie* 51 (1969), 260-285.
- Ashworth, E. J.: *Language and logic in the post-medieval period*, Reidel, Dordrecht, 1974.
- Ashworth, E. J.: "Chimeras and imaginary objects: a study in the post-medieval theory of signification", *Vivarium* 15 (1977), 57-79.
- Boehner, Philotheus: *Medieval logic. An outline of its development from 1250 to c. 1400*, University Press, Manchester, 1966.
- Cassin, Chrystine E.: "Russell's discussion of meaning and denotation: a re-examination", *Essays on Bertrand Russell* (E. D. Klemke, ed.), University of Illinois Press, Urbana, 1970, 256-272.
- Ducrot, Oswald: "La théorie médiévale de la supposition", *History of linguistic thought and contemporary linguistics* (H. Parret, ed.), de Gruyter, Berlin, 1976, 189-277.
- Ebbesen, Sten: "The chimera's diary", *The logic of being* (S. Knuuttila and J. Hintikka, eds.), Reidel, Dordrecht, 1986, 115-143.
- Geach, Peter Thomas: *Reference and generality. An examination of some medieval and modern theories*, Cornell University Press, Ithaca, 1980 (tercera edición).
- Griffin, Nicholas, "Denoting concepts in *The principles of mathematics*", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 23-64.
- Henry, Desmond Paul: "The early history of suppositio", *Franciscan studies* 23 (1963), 205-212.
- Hursthouse, Rosalind: "Denoting in 'The principles of mathematics'", *Synthese* 45 (1980), 33-42.
- Hylton, Peter: *Russell, idealism and the emergence of analytic philosophy*, Clarendon Press,

³⁰ Ver E. Karger, "Consequences et inconsequences de la supposition vide dans la logique d'Ockham", 52.

³¹ Karger, por ejemplo, declara haber encontrado en Ockham *dos lógicas* distintas, una aplicable a proposiciones con sujeto vacío y otra no.

- Oxford, 1990.
- Hylton, Peter: "Beginning with analysis", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 183-216.
- Karger, Elizabeth: "Consequences et inconsequences de la supposition vide dans la logique d'Ockham", *Vivarium* 16 (1978), 46-55.
- Kretzmann, Norman: "Medieval logicians on the meaning of the *propositio*", *The journal of philosophy* 67 (1970), 767-787.
- Matthews, Gareth B: "*Suppositio* and quantification in Ockham", *Noûs* 7 (1973), 13-24.
- Moody, Ernest A.: "The medieval contribution to logic", *Studies in medieval philosophy, science and logic*, University of California Press, Berkeley, 1975, 371-392.
- Noonan, Harold: "The 'Gray's Elegy' argument —and others", *Bertrand Russell and the origins of analytical philosophy* (R. Monk y A. Palmer, eds.), Thoemmes, Bristol, 1996, 65-102.
- Nubiola, Jaime: *La renovación pragmatista de la filosofía analítica. Una introducción a la filosofía contemporánea del lenguaje*, Eunsá, Pamplona, 1994.
- Pérez-Ilzarbe, Paloma: "The doctrine of descent in Jerónimo Pardo: meaning, inference, truth", *Studies on the history of logic. Proceedings of the III. Symposium on the History of Logic* (I. Angelelli y M. Cerezo, eds.), de Gruyter, Berlin, 1996, 173-186.
- Pérez-Ilzarbe, Paloma: *El significado de las proposiciones. Jerónimo Pardo y las teorías medievales de la proposición*, Eunsá, Pamplona, 1999.
- Pinborg, Jan: *Logica e semantica nel medioevo*, Boringhieri, Torino, 1984.
- Rijk, Lambert Marie de: *Logica modernorum. A contribution to the early terminist logic*, Van Gorcum, Assen, 1962, 1967.
- Russell, Bertrand: *The principles of mathematics*, George Allen & Unwin Ltd, London, 1972. (citado como PoM)

Paloma Pérez-Ilzarbe
Universidad de Navarra
e-mail: pilzarbe@unav.es



FREGE AND HEIDEGGER ON REFERENCE DETERMINATION AND SENSE*

DANIEL QUESADA

Universitat Autònoma de Barcelona

ABSTRACT: Frege has a view on the relation of sense and reference to which the possibility of sense without reference is integral (sense determines reference *intentionally*). This view is a semantic underpinning to traditional representationalism, and can be challenged in two opposite directions. One is a thorough (externalistic) realism, and the other accentuates internalism. A view of sense in this later direction is first roughly outlined (but not endorsed). It is then argued, using some textual evidence, that it is indeed Heidegger's view of sense, resulting in an *ontological determination* of reference, and in much which is central to his philosophy.

KEYWORDS: Sense, reference determination, representationalism, internalism, externalism, Frege, Heidegger.

Introduction

It was about 100 years ago that Gottlob Frege introduced the distinction between *sense* (*Sinn*) and *signification* or *reference* (*Bedeutung*):

It is natural (...) to think of there being connected with a sign (name, combination of words, written mark), besides that which the sign designates, which may be called the signification of the sign, also what I should like to call the *sense* of the sign, wherein the mode of presentation is contained (...) The signification of 'evening star' would be the same as that of 'morning star', but not the sense [Frege (1892), p. 57]¹.

* This paper is part of research for project PB97-0192 funded by the DGES, Spanish Ministry of Education, and the "joint action" project HB97-0147, funded by the DGES and the British Council.

¹ I substitute the word 'signification' for the misleading 'meaning' which the authors of the translation use to render the German word '*Bedeutung*'. In general, I will use the pair 'signify' and 'signification' and also the more technical 'refer' and 'reference'/'referent'—which is likely worst but common—for the German words 'bedeuten' and 'Bedeutung' respectively.

Thirty five years later, in a work that was to become world famous, Martin Heidegger explained a notion of *sense* of which the following is a part:

Sense (Sinn) is that wherein the understandability of something holds up [Heidegger (1927) p. 151]².

Is there any interesting connection between Frege's notion of *Sinn* and Heidegger's notion, or is it a case of an ordinary word which is given two basically unrelated, or at most loosely related, technical meanings? Certainly, it is not far-fetched to conjecture some kind of connection which passes through Husserl. As it is well known, Husserl notion(s) of *Sinn* in his *Logical Investigations* is (are) related to Frege's (it seems that not only conceptually, but historically)³, and there is evidence that Heidegger knew about this aspect of the philosophy of his mentor Husserl. But is it there a connection which is philosophically interesting? A positive answer to this question can be of some interest, especially if it were to throw an additional sidelight on discussions around the Fregean notion, central as these are to current research in philosophy. But even if it did not, it has the potential interest of clarifying the links between the ideas of two of the major forces configuring the main broad division present in contemporary philosophy: that between analytical and continental philosophy. Both philosophers have been made to play central roles in the respective camps, and their ideas on *Sinn* are all but tangential to an influence that runs through the most part of the XXth century and, on all odds, into the next millenium.

The question of the connection has been (implicitly) answered in the affirmative in Lafont (1994), and what will be said here is in substantial agreement with her position, although the issue is given here a different twist, so to speak.

A "briefing" on Fregeanism and anti-Fregeanism

While otherwise not explicitly indicated, let us for simplicity restrict our discussion to "physical world discourse", that is, to the domain of language sentences which are normally used to intend to make claims about the physical world, and to utterances of such sentences. Thus restricted, the relation of reference or signification is the relation between (bits of) language and (parts of) the (physical) world or (physical) reality. Our discussion will be further limmited by taking into account only the sense and signification of singular terms appearing in sentences or utterances, that is to say, names, demonstratives, indexicals and complex demonstrative or indexical expressions

² All translations of Heidegger's texts are my own, although I have consulted other translations when available, in particular the English translation of *Sein und Zeit: Being and Time*, Oxford: Blackwell, 1962 (10th reprint 1992).

³ For the conceptual connections see Künne (1986), especially pp. 195-198.

built out of these (most easily by adding a common noun: 'this table', 'that door')⁴. Finally, I will proceed implicitly as if the sense of demonstrative and indexical expressions could be specified in the same way as the sense of proper names, which it cannot.

The Fregean notion of sense is that of the *epistemic value* of an expression for a speaker or a thinker. As it is usually explained, Frege's doctrine was that the sense of a name can be expressed by a pure definite description (a definite description containing only general terms) which is *synonymous* with the name. The reference or signification of the name is then that unique entity—if any—to which the (general) concepts corresponding to the terms in the description apply⁵.

According to this view, the sense of a name sticks to the name-type and only conceptual material constitutes it; objective entities or—as the restrictions in our discussion allow us to say—physical things are altogether excluded. And why are they excluded? The appealing intuitive idea is that sense being constitutively linked to the epistemic value of expressions, it is thereby connected with understanding, and one can, it would seem, understand what is said in uttering a sentence containing names without any inkling as to whether the names in the utterance refer to any (physical) object.

On such a view, the relation between the sense of a proper name and its signification is *extrinsic*, since it is contingent that such a sense has a signification: if extralinguistic reality were different, that *same name* (type) with that *same sense* could fail to have any referent. If and when there is an object signified, that object is the signification or referent *determined* by the sense. Let us talk of *intentional determination* to speak about this view on how senses relate to objects signified, on account of the central and direct way in which Frege's theory of the senses of proper names incorporates the trait of fallibility that characterizes intentional relations.

Now, it is widely accepted, due mainly to Saul Kripke's *Naming and Necessity*, that Frege's theory of sense-and-reference, in the way it has been formulated, is untenable. One reaction to this has been a return to the Millian doctrine that the only semantic property of a name is that of referring or

⁴ Besides linguistic utterances, we could include also *mental occurrences*—such as mental acts of perception—purportedly about bits of the physical world. We will do this below.

⁵ Kaplan exemplifies widespread opinion when he describes a Fregean sense as "a concept, something like a description in purely qualitative language"; the referent is "the individual who falls under the concept, i.e., who, uniquely, has the qualities", and the relation between sense and reference is labeled "in general empirical" [see Kaplan (1989), p. 485]. Of course the concept can be a complex concept or a complex of concepts (as it is normally the case with the ones associated to descriptions). Calling the relation 'in general empirical' is not very fortunate, since it is beside the point whether empirical information coming through the senses (of sight, etc.) plays a central role or not. 'Contingent' would be therefore better.

signifying a bearer; any “connotations” a name may have will belong in a psycholinguistic appendix, or perhaps —as it is added nowadays— they will be part of an explanation of the practice of using a name for the bearer it has in fact. In any case, it does not belong in the central core of a semantic theory which aims at delineating in a systematic way the conditions upon which the truth of utterances depend. When it comes to give the representational content of sentences and utterances (where this representational content includes physical objects), any considerations relevant to their *understanding*, which could perhaps be given by a theory of senses, is out of place.

However, Dummett and Searle have developed Fregean theories which 1) follow Frege’s doctrine in associating “purely descriptive” senses to names, but that 2) by a new careful selection of the kind of general concepts involved in senses succeed in “avoiding the counterexamples”, that is, succeed in accounting for Kripke’s intuitive data on possible world truth conditions for sentences. Moreover, Dummett strongly advocates a theory of meaning as a theory of understanding or a theory about what competent speakers know when they understand the meaning of expressions. Representational content that includes physical objects is excluded from such a theory:

Reference ... is not an ingredient in meaning at all: someone who does not know the reference of an expression does not show thereby that he does not understand, or only partially understands, the expression [Dummett 1973, p. 84].

... reference is not part of the meaning —it is not part of whatever is known by anyone who understands the expression [Dummett 1991, p. 123].

Thus, we have here the exact antithesis of the Millian alternative I mentioned above. On the face of it, however, there is a contention that *both* thesis and antithesis endorse, to wit, *that someone who does not know the reference of an expression does not show thereby that he does not understand, or only partially understands, the expression*. But the disputing parties draw exactly opposite conclusions out of this background agreement. “That is why we must *exclude references* from a theory of meaning *stricto senso*” says de Fregean-Dummettian. “That is why we must *exclude understanding* (or *knowing* of references) from a theory of meaning *strictu senso*”, claims the Millian.

A third party has recently joined the dispute. They want to deny precisely the contention common to the two other parties. However, this is no mean enterprise, since the intuition expressed thereby is not just any ordinary intuition. For agree that we were disposed to think that someone that utters a sentence or hears an utterance containing a name without a referent does thereby have only at most an imperfect understanding of the utterance. Then (flawless) understanding would depend on whether the names used have a

referent, and so we would not know for sure, nor immediately, whether we (fully) understand or not, since it is not upon us whether the names have a referent or not (at least where the referents are physical objects). But this, it seems, conflicts with the idea that we have a direct and certain knowledge of the contents of our thoughts, including the thoughts in the (perhaps potential) entertaining of which understanding an utterance consists. And *this* idea has run deep through the history of philosophy, since Descartes at the very least.

The possibility of a *theory* of meaning as understanding that *includes* references is not altogether foreclosed by the *intuition* that excludes reference from sense, or from a theory of meaning as understanding, since, with the proper theoretical motivation, one could regard cases of utterances including names without reference as derivative, as Evans and McDowell in their work on senses urge us to do. Nevertheless, in view of the close links between this intuition and the deep-rooted philosophical idea, to motivate and develop such a theory is bound to be no easy undertaking, as I said [but see recent promising developments in García-Carpintero (1998) and (forthcoming)].

Placing Heidegger's views on sense where they belong

In the next paragraphs I would like to delineate a possible view of sense and its relation to reference, and to relate this view with the positions we have reviewed above. Then I will argue, using some textual evidence, that the view fits central relevant pronouncements by Heidegger.

In order to motivate the view I want to describe, it will be convenient to start by quoting the following text by McDowell:

The Fregean view would have to seek its support in the idea that thought relates to objects with an essential indirectness (...) Whether the object exists or not would then be incidental to the availability of the thought. Underlying that idea is the following line of argument. When we mention an object in describing a thought we are giving only an extrinsic characterization of the thought (since the mention of the object takes us outside the subject's mind); but there must be an intrinsic characterization available (one which does not take us outside the subject's mind), and that characterization would have succeeded in specifying the essential core of the thought even if extra-mental reality had not obliged by containing the object [McDowell (1977), p. 153].

In thus describing the Fregean view, McDowell's rethorics suggests clearly the possibility of denying the coherence or tenability of a successful characterization of a thought independently of whether the reality "obliges" by containing the object. What I wish to point out is simply that there are *two*—exactly opposite—ways of denying this coherence (or tenability). One of them is to maintain that physical objects—or in general objective, mind-

independent objects— belong essentially to any characterization or individuation of a thought, or of the sense of a sentence or utterance. This is, undoubtedly, the one that McDowell has in mind. But there is another one, and this is that, since objects belong essentially in any characterization or individuation of a thought, and *physical* objects are definitely out of place in there, there are *other kind of objects* that belong in the characterization or individuation of a thought.

In agreement with what McDowell says, Frege's position is akin to classical Lockean representationalism, where physical objects are never part of what is "grasped by the mind", of what is understood. McDowell aimed-to position would be a fully-fledged realism which opposes this. And the third view—opposed also to Frege's representationalism—is the one that we want to describe closer now.

An obvious place to start is to ask: What did we mean by 'other kind of objects'? What or which are these objects? And the simple answer, to be developed below, is: "The objects to which the concepts constituting the senses of the names (or other singular terms) apply to, that is, the objects which, in any case of use of a singular term, uniquely have the properties corresponding to the concepts constituting the sense of that singular term." Now, what are these properties? According to the position I am outlining, these properties are *projected properties*, that is, properties entirely "constituted" by our decisions to apply or not to apply concepts. It is important to remark here that this 'our' must be understood as a collective or social 'our'. That is, it is not the case that an individual, by deciding to apply or not to apply a concept in the different circumstances that prompt the decision, "constitutes" a property. It is the "collectivity" that "constitutes" it, and this means the decisions that are considered as normal in a social group that speak an historical language. The idea, then, is the familiar one that any internally consistent combination of concepts corresponds to an object, which is the referent of the singular term whose sense consists just in that conceptual combination, only that, among such objects, we distinguish those which exist from those which do not (more about this in a moment). Thus, from this viewpoint, any name or singular term with a (coherent) sense—and this is redundant, since a conceptual combination which is not coherent does not form a sense—has a referent. To add now, as a side comment, that referents are thus regarded as "immanent to discourse" would suggest that it is legitimate to draw a distinction between what is immanent and what is transcendent to discourse, something that does not make any sense for the position outlined.

What is meant, according to this view, by saying that some (possible) objects exist while others do not? We must realize that, for the position being described, there is no more to truth than what is collectively considered the

right —accepted— thing to say in the circumstances at issue by the collectivity situated in an historical period speaking a certain language. Suppose then that, for example, ‘*a* does not exist’ is (regarded as) true, where ‘*a*’ is a singular term with a sense. This means that the entity which is the referent of ‘*a*’ —and there is necessarily such an entity, an entity exactly on a par with all other entities allowed by intelligible discourse— has whatever properties the collective of speakers (of a natural language in the historical period considered) associate to existence in the period at issue: maybe that we can see it, or maybe that we can “feel intuitively” its presence, what in turn —and this must not be overlooked— means that it would be accepted to utter things like ‘*a* can be seen’ or ‘*a* can be intuitively felt’. And whether that entity has those properties or not depends entirely on the *conception* of *a* current at the time and place, what amounts to say that it depends entirely on the sense of ‘*a*’, and so, ultimately on discourse.

As we see, from the viewpoint outlined, there is no more to objects that what belongs to them in virtue of our concepts —the senses of our discourse—, and this remark does not only concern the properties such objects have, but their very existence as well. Accordingly, it seems appropriate to allude to this whole view of the sense of expressions and its relation to the objects which are their referents by saying that, on this view, reference is *ontologically determined* by sense. Ontological determination is a way of conceiving the sense-reference relation clearly distinct from intensional determination, as is most easily seen by one of its central consequences, to wit, that there is no sense without reference. But it is nevertheless at the other extreme from the position sought for by the philosophers that are currently trying to articulate a notion of object-dependent senses, and which would nevertheless agree (rather nominally) to the “no sense without reference” claim. In the end, what we have here are two completely opposite ways of closing the gap —unbridgeable, given Frege’s representationalism— between the understanding of a sentence or utterance and its representational content or the condition for its truth (when this is conceived of as including the referents of the names or singular terms which appear in the sentence or utterance).

Next, I would like to make at least plausible that Heidegger maintained a view of sense and reference that corresponds with the view outlined above, the one which is epitomized by ontological determination. I will concentrate above all on the notion of sense, hoping that part of the remaining at least can be glimpsed by the reader. In any case, a fuller justification of the claims which will be made or suggested would require much more space than I have here.

Heidegger is explicit in endorsing the every-name-a-referent principle in his postdoctoral thesis (thesis for entitlement to teaching) on Duns Scotus, where he paraphrases aprovingly Scotus’s “wide conception of the way of signifying of names” according to which “[a] name signifies (*bedeutet*) its

object, whether it exists or it does not exist" [Heidegger (1915), pp. 348-349]. The reason he gave for this approval is —again, it seems, in agreement with Scotus— that "the way of signifying of names gets its content (*Sinn*)" from "the most general determinations that hold for objects (*Gegenstände*)", and these consist in being able to be the objects (*Objekte*) for the [intentional] acts directed towards them". On the face of it, a rather Brentanian reason, according to which *any* intentional act has an object, which is immanent to the act.

Of course, there is no trace in this passage —nor indeed is there, to my knowledge, in the whole work from which it is taken— of any concept of sense bearing a close relationship to the Fregean concept (the word 'Sinn' appears rather anecdotically in the passage, so to speak, which is reflected in the non-literal translation chosen above). What is endorsed here is only the every-name-a-referent principle, which is implied but does not imply the principle we are after. So far as I know, however, Heidegger never disowned that principle.

Heidegger's scanty remarks on a notion of sense in *Being and Time* appear in § 32 ("Understanding and Interpretation") and are reviewed and expanded at the beginning of § 65 ("Temporality and the Ontological Sense of Care"). This comes close to an attempted definition:

Sense is the structured "upon-which" of the projection in terms of which something becomes intelligible as something [Heidegger (1927) p. 151, italics in the original].

To get a first approximation to an understanding of what Heidegger is getting at here and also in the text quoted at the beginning of this paper, I will organize the explanation around a simple sentence (of the kind Heidegger himself uses for his examples): 'This table is black'. Now, the notion of projection, that is involved in the text just quoted, is not having an straightforward explanation, not least because Heidegger plays with the various connotations of the word (the German 'Entwurf'). Heidegger talks of understanding as "projecting" human beings (actually: Dasein's Being) "upon [their] 'for-the-sake-of-which'" (*op. cit.* p. 145). This 'for-the-sake-of-which' refers to the domain of "teleological" properties linked to ordinary life which for Heidegger play a privileged role, but whose more exact determination does not concern us here. They are properties like "for writing", "for going in and out", "for lighting up", "for sitting" [see Heidegger (1925-26), p. 144]; as said, determining the exact domain of the basic properties does not concern us now. The next thing to notice is that the concepts corresponding to these properties can be combined, with the only limitation of conceptual consistency, giving then rise to a conceptual structure that points to possibilities in that there can be objects to which those concepts apply. This structure is, I believe, what is called by Heidegger the 'as-structure' (*Als-Struktur*): all objects that can be perceived, or, in general, all objects of any intentional acts, in so far these acts

are indeed carried out, are necessarily immediatly understood in terms of this structure (cf. *loc. cit.*).

Because sense is *that upon which* understanding is “projected”, the word ‘Sinn’, as used technically by Heidegger, refers to the whole structure of concepts I have (very roughly) described, the concepts in which the “as-structure” consists. Heidegger is thereby using the word ‘Sinn’ to cover the *collection* of what we would call, using the Fregean terminology, ‘senses’, or perhaps more accurately, a subcollection of them. How is this claim justified? The only possible rival interpretation I can think of is that the components of the domain of the “as-structure” are the semantic values of predicates. Well, as a matter of fact they are, but, according to Heidegger, not primarily⁶:

The pre-predicative character of the as-structure is to be explicitly emphasized because one could think, relying naturally on the linguistic expression, that this as-structure is first and properly given in the simple statement: “This table is black” (...) but (...) it is to be understood that this “as” is not really primary of predication *qua* predication, but it is previous to it, so that it makes the structure of predication possible. Predication has the as-structure, but in a derived way (...) the as-structure does not belong to what is thematically grasped; it can indeed be grasped, but not as the first in the row and in the direction of grasping a table, a chair and so on (*op. cit.* pp. 144-145).

There is a second related technical use of ‘Sinn’ by Heidegger in which the word applies to *entities* (for this use see especially *op. cit.* § 65), especially to abstract entities, such as, for example, “care”. But I think that this use is readily understood from the first when one takes into account also the Fregean use according to which ‘Sinn’ covers de “way of presentation” of the *entities* which are the referents of expressions.

Moving on now finally to the notion of reference or signification, it is indeed true that the technical notion that Heidegger associates with ‘bedeuten’ in *Being and Time* and theoretically related works is quite unrelated to the Fregean notion, or, indeed, to the notion Heidegger himself used in 1915 in the work mentioned above, which are in any case much closer to common German usage. Nevertheless, the claim that, according to Heidegger’s views, significations or references (in the Fregean sense of the word) depend ontologically on senses would be justified by laying out some of the numerous texts in which Heidegger says that understanding—the concepts of the as-structure above all—is primary and necessary to any cognitive access to

⁶ These claims about the ‘as-structure’ are emphasized in Lafont (1994). See especially § 1.4.2.2 in which she polemizes with E. Tugendhat regarding the character of the “pre-predicative” in Heidegger.

objects. Indeed, having the later Heidegger transferred the *locus* of understanding from Dasein to language, I think that pronouncements like the following, which may well startle us for their extremism, apply also to the relation between expressions —*any* denoting expressions with a sense, and just because they have a sense— and their referents:

Only when the word for the thing is found, is the thing a thing. So, only then *is*. (...) Only the word provides its being to the thing [Heidegger (1959), p. 164].

REFERENCES

- Dummett, Michael: *Frege: Philosophy of Language*, London: Duckworth, 1973, 2nd. edition, 1981.
- Dummett, Michael: *The Logical Basis of Metaphysics*, Cambridge, Mass.: Harvard University Press, 1991.
- Frege, Gottlob: "On Sense and Meaning", in Geach, Peter. and Black, Max. (eds.), *Translations from the Philosophical Writings of Gottlob Frege*, third edition, Oxford: Blackwell, 1980 (original 1892).
- García-Carpintero, Manuel: "Indexicals as Token-Reflexives", *Mind* 107 (1998), 529-563.
- García-Carpintero, Manuel: "Senses as Presuppositions of Acquaintance", *Journal of Philosophy* (forthcoming).
- Heidegger, Martin: "Die Kategorien- und Bedeutungslehre des Duns Scotus", in *Frühe Schriften* (1912-1916), *Complete Works (Gesamtausgabe)*, vol. 1, Frankfurt a. M.: 1978 (original 1915).
- Heidegger, Martin: "Logik. Die Frage nach der Wahrheit", in *Complete Works (Gesamtausgabe)*, vol. 21, Frankfurt a. M.: 1976 (original 1925-26).
- Heidegger, Martin: *Sein und Zeit*, 16th edition, Tübingen: Max Niemeyer, 1927.
- Heidegger, Martin: *Unterwegs zur Sprache*, 8th edition, Pfullingen: Neske, 1986.
- Kaplan, David: "Demonstratives", in Almog, Joseph, Perry, John and Wettstein, Howard (eds.), *Themes from Kaplan*, Oxford: Oxford University Press, 1989.
- Künne, Wolfgang: "E. Husserl: Intentionalität", in Speck, J. (ed.), *Grundprobleme der großen Philosophen: Philosophie der Neuzeit IV*, Göttingen: Vandenhoeck, 1986.
- Lafont, Cristina: *Sprache und Welterschließung*, Frankfurt: Suhrkamp, 1994.
- McDowell, John: "On the Sense and Reference of a Proper Name", in Platts, M. (ed.) *Reference, Truth and Reality*, London: Routledge and Kegan Paul, 1980 (original 1977).

Daniel Quesada Casajuana
 Universitat Autònoma de Barcelona
 e-mail: Daniel.Quesada@cc.uab.es

LA FILOSOFÍA ANALÍTICA Y LA TRADICIÓN ANALÍTICA: Un siglo a vueltas con el significado

M^a UXÍA RIVAS MONROY
Universidad de Santiago de Compostela

ABSTRACT: The question I want to consider in this paper is to what extent Analytic Philosophy is alive nowadays. My answer to that question will consist in establishing a distinction between Analytic Philosophy and the Analytic Tradition. As many other people I also think that Analytic Philosophy was a way of thinking which flourished in the first part of our century. The Analytic Tradition, like a new generation, has new features without losing the family resemblances which link it with Analytic Philosophy. What I will try to do here is to establish similarities and differences between this two ways of understanding philosophy -and mainly, philosophy of language- which have so much in common.

KEYWORDS: Analytic philosophy, analytic tradition, philosophy of language, meaning.

1. Introducción

En estos momentos en los que el próximo milenio está cada vez más cerca, y con la perspectiva que da el tiempo, podemos intentar ofrecer una brevísima caracterización de la corriente de pensamiento denominada Filosofía Analítica, que tuvo su época de mayor esplendor en la primera mitad de este siglo. La Filosofía Analítica dio lugar a lo que ahora llamamos el giro lingüístico, al afirmar, de muy diferentes maneras, que una gran parte de los problemas filosóficos eran en realidad problemas lingüísticos que el análisis -principalmente lógico, pero también lingüístico- podría resolver o incluso disolver. Frege, Russell, Wittgenstein son las figuras más representativas de la defensa de esta idea, pero también la Filosofía del Lenguaje Ordinario, o el Neopositivismo Lógico. Después de 1950, y sobre todo en Estados Unidos, podemos decir que otros filósofos siguen las pautas marcadas por este movimiento, aunque ya no defiendan explícitamente la idea de que los problemas filosóficos son sobre todo problemas lingüísticos que el análisis

solucione. Filósofos como Quine, Davidson, Dummett, Searle, Kripke, Putnam o Burge, entre otros, son los herederos de la Filosofía Analítica, y en este sentido yo diría que pertenecen a la Tradición Analítica, sin identificarlos plenamente como filósofos analíticos. Quizás con más perspectiva histórica lo que ahora denominamos Filosofía Analítica acabe incluyéndolos a ellos también, pero, como indicaré a lo largo de estas páginas, ciertas maneras de abordar los temas relacionados con el significado, la referencia, la verdad o la fuerza de los enunciados, me llevan a situarlos bajo el rótulo de la Tradición Analítica. Aglutino a estos autores bajo la denominación de la Tradición Analítica porque la mayor parte de las discusiones filosóficas sobre el significado y temas afines las realizan teniendo como telón de fondo la Filosofía Analítica. Estos autores vuelven a ocuparse de los mismos -o muy semejantes- problemas de los que en su momento se ocuparon los filósofos analíticos, pero ellos aportan otro tipo de soluciones, introduciendo nuevos elementos tales como el holismo, el pragmatismo, el contexto y ciertos aspectos sociales y mentales, que quedaban fuera de las consideraciones de los autores que les precedieron.

Centraré mis comentarios sobre la filosofía del lenguaje y el significado. Las diferentes clases de reificación de los significados, la reacción contra el psicologismo o la metafísica, la importancia de la lógica y de los lenguajes formales son algunos de los pilares que dieron forma a la Filosofía Analítica, de la que podemos decir que abordó el análisis del lenguaje desde una perspectiva doblemente externalista. Por un lado, al estudiar el lenguaje en sí mismo, atendiendo a sus aspectos sintácticos; y, por otro lado, al prestar atención a su dimensión semántica, al estudiar el lenguaje en su relación con el mundo, destacando casi exclusivamente en esta última relación su función descriptiva. La Tradición Analítica rechaza los significados como entidades, pone el acento en la construcción de teorías del significado y da importancia a los elementos pragmáticos, tales como el contexto, la comunicación y la sociedad. La Tradición Analítica enfatiza también la perspectiva internalista del lenguaje tomando en consideración aspectos psicológicos, que acercan la filosofía del lenguaje a la filosofía de la mente, y que suponen una reconsideración del giro cartesiano. Pero especialmente son los aspectos pragmáticos los que esta tradición más ha tenido en cuenta, mostrando el vínculo tan estrecho que existe entre el lenguaje y los hablantes, incidiendo una vez más en que el lenguaje es ante todo un hecho social.

2. Primer movimiento. La filosofía analítica: descubriendo la cara oculta del lenguaje

Aunque en la Filosofía Analítica los aspectos sintácticos del lenguaje fueron objeto de una gran atención, y llegaron a su punto culminante con la concepción carnapiana de la filosofía como sintaxis lógica, los aspectos semánticos del lenguaje no fueron sin embargo menos tenidos en cuenta, especialmente en lo que concierne al significado y a la referencia, si bien también aquí predominó una cierta perspectiva externalista al propio lenguaje.

Suele situarse el origen de la Filosofía Analítica en la figura de Frege. Frege es importante por muchas razones, pero en el contexto del tema que nos ocupa, es importante, a mi modo de ver, porque él marcó, o indicó, ciertas líneas o cauces por los que después discurrió la Filosofía Analítica. En primer lugar, estableció dos relaciones semánticas fundamentales de las palabras y de las oraciones: el sentido y la referencia. Estas relaciones semánticas son de distinto carácter, pero Frege las consideraba como la relación entre un instrumento, el lenguaje, y algo externo al mismo: el mundo de los pensamientos, en un caso, y el mundo de los objetos, en el otro. Así el sentido era la relación entre el lenguaje y los pensamientos que éste expresaba, y la referencia era la relación entre el lenguaje y los objetos no lingüísticos que éste denotaba. Esto era lo que ocurría en el discurso ordinario. En el discurso indirecto los pensamientos eran también los objetos referidos por las oraciones del lenguaje. La distinción entre sentido y referencia en las expresiones resolvía problemas filosóficamente interesantes, tales como la explicación de la diferencia existente entre enunciados de identidad analíticos y enunciados de identidad sintéticos. También aportaba una explicación al distinto comportamiento, con respecto a su valor de verdad, que tiene una oración si ésta aparece en un contexto de actitud proposicional. Resolvía igualmente, aunque con muchas dificultades, el hecho obvio de que hay expresiones que no se refieren a nada existente y no por ello carecen de significado.

La distinción sentido/referencia fue cuestionada por Russell especialmente en su aplicación a los nombres propios, pero Russell tampoco aceptó la principal afirmación de Frege de que la referencia de un enunciado era su valor veritativo. A partir de aquí la opción por una u otra postura con sus respectivas argumentaciones fue moneda de cambio en los años siguientes, e incluso en la actualidad. Las recientes discusiones acerca de si los nombres propios denotan directamente, o si lo hacen a través del sentido a ellos asociado; o si las oraciones se refieren a un valor veritativo o a una proposición singular tienen su origen en esta primera confrontación de posturas, y hoy en día es corriente hablar de aquellos que defienden la posición descriptivista y la

posición de la referencia directa bajo el nombre de neo-fregeanos y neo-russellianos.

Otro de los puntos en los que el pensamiento de Frege abrió camino fue en la actitud de permanente sospecha con respecto al lenguaje natural, por ser defectuoso, impreciso, ambiguo, en resumen, por ser poco apropiado para la expresión del conocimiento científico. En este sentido la lógica sería el instrumento que clarificaría la expresión de los resultados científicos, dotando de rigor y fiabilidad al lenguaje de la investigación y, por tanto, a la ciencia. La estrecha relación entre lenguaje, lógica y ciencia aparece claramente en los escritos de Frege, donde la Conceptografía -una nueva lógica alejada de la gramática y próxima en su formulación a los lenguajes formales matemáticos-aplicada a los lenguajes naturales permitiría el control de calidad exigible a la expresión del conocimiento científico. Russell y Wittgenstein, de formas diferentes llevaron a cabo esta aplicación que sugería Frege. Russell resolvía interesantes problemas de filosofía del lenguaje con lo que él consideraba uno de sus mayores logros: su teoría de las descripciones definidas. Wittgenstein dotaba de dimensión filosófica a la estructura lógica del lenguaje, al establecer la lógica como el límite de lo que se puede decir y de lo que se puede pensar y, por consiguiente, como el límite de la realidad y del mundo.

Siguiendo a Wittgenstein en *Las Investigaciones filosóficas*, el error cometido por Frege, Russell o él mismo al insistir en la búsqueda de la forma lógica del lenguaje y de su estructura, que metafísicamente considerada reflejaría la estructura del mundo, fue querer encontrar "la esencia", "lo profundo", "lo oculto". Quizás la forma lógica, apuntada como condición de posibilidad del mundo y de nuestro pensar sobre el mismo, fue un gran acierto del *Tractatus* de Wittgenstein, que aún hoy en día nos inquieta y estimula. Esta investigación sobre la forma lógica del lenguaje, y su aplicación a los lenguajes naturales para resolver ciertos problemas semánticos, estuvo directamente vinculada al estudio de los fundamentos de la lógica y de las matemáticas, a la clarificación de muchas nociones lógicas, y al prodigioso avance de la lógica, de las matemáticas, y de los lenguajes formales en general. Y estos avances posibilitaron la definición de una noción semántica tan compleja como evidente resulta su intuición, que hasta entonces fue tomada en algún caso como una noción primitiva, y por ello indefinible; me refiero a la noción de verdad, definida por primera vez por Tarski en 1933.

Otra importante contribución de Frege y que, posteriormente desarrollada, es una de las marcas características de la Filosofía Analítica es el situar al enunciado como unidad básica del significado y, junto con ello, destacar como función primordial del lenguaje la función descriptiva. El

convencimiento y la defensa de estas ideas tuvo interesantes consecuencias ontológicas y epistemológicas, primero bajo la forma del atomismo lógico, y, posteriormente, llevadas al extremo, dando lugar a la rama más conservadora de la Filosofía Analítica: el Neopositivismo.

Si el punto de partida lo situamos en Frege, y en su afirmación de que el significado de una palabra hay que buscarlo en el contexto de la oración, y de que la verdad o la falsedad se dice de los sentidos o pensamientos expresados por enunciados; el punto de llegada hay que situarlo en los Positivistas Lógicos, en su defensa del principio de verificación y de la distinción entre verdades analíticas y verdades sintéticas. Lo interesante de esta distinción, tan antigua casi como la propia lógica, no es la distinción en sí misma, sino la argumentación empleada para defender que éstos son los únicos enunciados con sentido, y por ello, científicos, mientras que todos los demás son sinsentidos, es decir, pseudoenunciados. En esta conclusión juega un importante papel la concepción del significado defendida en un primer momento por los empiristas lógicos, y conocida como principio de verificación. Este principio de verificación se formuló teniendo muy en cuenta las aportaciones de la epistemología de Russell y las del *Tractatus*. La clara delimitación que hacía Wittgenstein en este libro entre proposiciones que describían los estados de cosas del mundo o de la realidad, y pseudoproposiciones fue fundamental para los neopositivistas. El conjunto de proposiciones verdaderas formaba el corpus de las ciencias. Las pseudoproposiciones eran para Wittgenstein también muy importantes, y cumplían otro tipo de funciones tan interesantes o más -para él seguramente mucho más interesantes- que la función descriptiva propia de las proposiciones, pero las pseudoproposiciones estaban situadas a otro nivel, distinto del que era propio del conocimiento científico, sin opción alguna a pasar por el filtro de la objetividad y la verdad; mostraban la parte más vital, mística e inexpressable del ser humano. Este aspecto de las pseudoproposiciones que Wittgenstein tanto valoró, fue completamente despreciado por los neopositivistas, que entendieron las pseudoproposiciones en el sentido más negativo posible equiparándolas a cháchara sin sentido e incluso sin significado. Pero como no era posible desacreditar de manera tan drástica a todo el lenguaje, hubo que establecer matices y diferenciar entre la función descriptiva y la función expresiva del lenguaje. Sabia decisión que no hizo posible sin embargo que la metafísica se salvara y pudiera tener algún hueco, aunque minúsculo, en el conjunto de los saberes, o incluso entre las artes. Wittgenstein concibió las tautologías como enunciados no descriptivos de estados de cosas que no dicen nada acerca de la realidad y que, por lo tanto,

no necesitan ningún método de verificación para determinar su significatividad, siendo, pues, siempre verdaderos en virtud de su forma lógica. Ello hizo posible que los neopositivistas evitaran la inclusión de estos enunciados en el reino del sinsentido al que el principio de verificación arrojaba tantas oraciones del lenguaje.

La epistemología de Russell, y del empirismo en general, también fue aceptada por los neopositivistas, y recogida en el principio de verificación, pero estas ideas fueron igualmente llevadas al extremo, y originaron un reduccionismo en dos direcciones. Una, la reducción del lenguaje de la ciencia a un lenguaje fenomenalista -que debido a los problemas de solipsismo a los que conducía, llevó más tarde a sustituirlo por un lenguaje fisicalista. Este reduccionismo iba ligado al proyecto de la unificación de la ciencia. Carnap trabajó duramente en la consecución de un lenguaje compatible con estas ideas. La otra reducción, que según Quine servía de base al dogma de la distinción entre enunciados analíticos y sintéticos, consistía en pensar que la base de la ciencia estaba formada por los enunciados protocolares o básicos, que se confirmaban uno a uno, y aisladamente considerados. Una consecuencia de este reduccionismo -junto con otras ideas- fue situar la filosofía únicamente del lado de las ciencias, separándola claramente de la metafísica. El culmen del giro lingüístico, y de la conexión entre lenguaje, ciencia, lógica y filosofía fue, entonces, considerar la filosofía como la sintaxis lógica de los lenguajes científicos. Como vemos, pues, las líneas sugeridas por Frege, desarrolladas y ampliadas de manera peculiar por Russell y Wittgenstein, obtuvieron su formulación definitiva y radical en el Positivismo Lógico, y especialmente en el Círculo de Viena.

De todo lo dicho es fácil ver como la Filosofía Analítica estudió e investigó el lenguaje desde una perspectiva externalista. La concepción del lenguaje como un medio de comunicación y de comprensión de nosotros mismos, de nuestros semejantes y de la realidad social que construimos fue obviada por completo. Los aspectos relacionados con la interiorización del lenguaje por parte de sus usuarios, los de nuestra comprensión del lenguaje como hablantes del mismo, la representación mental que nos permite expresar pensamientos propios y ajenos, el uso del lenguaje para inventar y novelar, no fueron tenidos en cuenta. Este interés por los aspectos más externos del lenguaje estuvo en la Filosofía Analítica directamente vinculado con el énfasis puesto en la función descriptiva del lenguaje, que llevó a entender el significado de las oraciones como la expresión de sus condiciones de verdad. Esto último estaba también relacionado con la defensa de compromisos ontológicos realistas, que llevaban a ver el lenguaje como un medio de representación de

una realidad objetiva que la ciencia iba descubriendo y desvelando; la radicalización de esta idea propició la concepción del lenguaje como una representación isomórfica del mundo, donde lenguaje y mundo comparten la misma estructura. Pero esta perspectiva abría también tímidamente una puerta al mezclar ontología y lenguaje; esta puerta la entreabrió Carnap al subordinar las cuestiones ontológicas al marco lingüístico, y la abrió de par en par Quine al introducir en ella la relatividad ontológica y la indeterminación de la referencia.

3. Segundo movimiento. La tradición analítica: la historia, las cosas y la sociedad también cuentan en el significado

La Filosofía Analítica también incluye la escuela denominada Filosofía del Lenguaje Ordinario, que surgió a la sombra de *Las investigaciones filosóficas* de Wittgenstein. Tanto esta obra como las contribuciones de los filósofos del lenguaje ordinario posibilitaron otra perspectiva del giro lingüístico: los problemas filosóficos son problemas lingüísticos porque los filósofos se alejan del uso ordinario de las palabras, y emplean las palabras descontextualizadamente, a la búsqueda de la esencia y lo profundo (un ejemplo de lo primero es el concepto de "nombre propio" en Russell, y de lo segundo, la teoría de la figura de Wittgenstein). La contribución más importante de este nuevo enfoque consiste en vincular los significados al uso, y los usos a las prácticas y a las actividades humanas. A través de Wittgenstein, Austin y Searle, entre otros, el lenguaje ya no es únicamente un medio que nos permite conocer la realidad, sino que con el lenguaje se hace realidad, puesto que con el lenguaje se realizan acciones, múltiples y variadas, y no simplemente se constatan hechos. Lenguaje, vida y acción son indisolubles. Acero ya le dio un nombre a esta manera de entender el lenguaje: "la concepción antropológica del lenguaje".

Esta concepción dejará su huella e influirá notablemente en la Tradición Analítica, a pesar de que ciertos autores siguen en varios sentidos siendo fieles a las pautas marcadas por el núcleo duro de la Filosofía Analítica: Frege, Russell y el Wittgenstein del *Tractatus*. Las soluciones aportadas por estos autores al abordar el significado dejaron en la tradición subsiguiente una gran insatisfacción, que llevó a otros, como Quine o Davidson, a optar por una vía conductista o extensionalista, fuera de todo tipo de compromisos ontológicos con respecto al significado. Si se acepta que el significado es equivalente a intensiones, sentidos, proposiciones, etc. parece que se está aceptando un reino de entidades abstractas con vida independiente, oscuras y difíciles de definir. En este sentido, la Tradición Analítica se caracteriza por un fuerte rechazo a cualquier tipo de reificación del significado, y, por lo general, califica

el significado, tanto de la oración como de las expresiones que la forman, como una abstracción metodológicamente útil para explicar la comprensión del fenómeno lingüístico. El significado así entendido se obtiene teniendo en cuenta igualmente el significado total de la oración y la contribución particular que cada palabra realiza a las distintas oraciones en las que se encuentra. Esta postura reacia a la reificación de los significados se encuentra también en *Las investigaciones filosóficas* de Wittgenstein, donde los significados se substituyen por usos, y desaparecen como entidades, reducidos al seguimiento de una regla; lo cual hace del lenguaje una actividad social y reglada, como muchas otras actividades sociales o institucionales.

La otra posibilidad de enfrentarse al significado sin caer en reificaciones es intentar reducir el significado a conceptos extensionales, como verdad, referencia, designación, denominación; pero las soluciones propuestas en este sentido, como la de Davidson, parecen perder algo de lo que intuitivamente entendemos por significado y que se plasma en las definiciones del diccionario. Por eso, estos programas extensionalistas se completan con la consideración de aspectos pragmáticos y empíricos, algunos ya insinuados en la Filosofía Analítica. Por ejemplo, la evidencia empírica procedente de los hablantes, el principio de caridad, o la incorporación de elementos pragmáticos en las teorías del significado, tales como la teoría del modo o la teoría de la fuerza.

Creo que una de las consecuencias de esta aversión al reificacionismo fue el interés por la elaboración de teorías del significado. La preocupación por las teorías del significado es algo nuevo que evidencia la complejidad que encierra la noción de significado. Éste se diversifica y disgrega; como en el caso de Dummett, que divide la teoría del significado en teoría de la referencia, teoría del sentido y teoría de la fuerza. Todos estos elementos -referencia, sentido y fuerza- configuran el significado. A las teorías del significado se les suele exigir una serie de requisitos que deben cumplir para ser teorías satisfactorias. En la dilucidación de estos requisitos se muestran ya las directrices que guían la concepción del significado que se propone, y que suele estar relacionada con la comprensión que los hablantes tienen de su lengua, pero también con el respeto a ciertos principios básicos heredados de la Filosofía Analítica, como el principio de composicionalidad, o la primacía de la función descriptiva del lenguaje sobre otras posibles funciones. Davidson, por ejemplo, es uno de los más fieles seguidores de una de estas ideas, al colocar como eje de su teoría del significado el concepto de verdad. Con ello deja entrever que la función más interesante del lenguaje sigue siendo la descriptiva, que se refleja en su afirmación de que una manera de conocer el significado de una oración es conocer sus condiciones de verdad. Así el significado se relaciona directamente

con el núcleo informativo-descriptivo de la oración. El programa de Davidson tiene entre sus objetivos extender esta concepción del significado al conjunto de todas las oraciones del lenguaje, analizando oraciones imperativas o desiderativas como si de enunciados se tratara. Sin embargo, ya el simple hecho de ocuparse de este tipo de oraciones, como de aquellas otras dependientes del contexto, muestra que las dos orientaciones que caracterizan a la Filosofía Analítica son tenidas en cuenta en la Tradición Analítica. Davidson puede servirnos otra vez de ejemplo, pues siendo quizás el más firme defensor de alguno de los principios básicos de la Filosofía Analítica en su orientación logicista, muestra una sensibilidad diferente a la que era propia de esta orientación en la Filosofía Analítica, al incorporar a su reflexión sobre el lenguaje la dimensión pragmática.

Aunque la Tradición Analítica se opone a las reificaciones del significado, no por ello tiene un espíritu tan combativo contra la metafísica como el que fue propio del Positivismo Lógico. Quine es el primero que da un paso a su favor, al desdibujar la frontera entre ciencia y metafísica. Hoy en día son muchas las cuestiones que podríamos calificar de metafísicas y que son abordadas por los filósofos de la Tradición Analítica: el esencialismo de Kripke, el relativismo ontológico de Quine, el realismo interno de Putnam. Desde esta óptica, la Tradición Analítica se enfrenta al reto de compatibilizar, en la medida de lo posible, realismo y relativismo. También con las reflexiones de estos autores, se pasa de la semántica realista propia de la visión cientifista del mundo y característica de la Filosofía Analítica a tener en cuenta también las semánticas posibilistas, favorecidas por la cantidad de experimentos mentales que ponen en tela de juicio las concepciones más clásicas del significado.

Entre estas concepciones clásicas, y su crítica consiguiente, destaca la presentada por Putnam. Las concepciones clásicas del significado son aquellas que incluso se pueden remontar a Aristóteles o Locke, donde el significado es entendido como una representación mental subjetiva, pero posiblemente compartida por el conjunto de hablantes de una lengua. Este modo de entender el significado fue ajeno en parte a la Filosofía Analítica -recordemos el antipsicologismo de Frege-, pero no completamente -pensemos en la importancia de los datos sensibles como referentes de los nombres propios lógicos de Russell, o como base de la verificación empírica en los neopositivistas, o, transformados en estimulaciones neuronales o irritaciones sensoriales, como configuradores del significado estimulativo en Quine. Putnam arremete contra la tesis de que el significado entendido como representación mental subjetiva pueda ser compatible con uno de los

principios fundamentales de la Filosofía Analítica, a saber, la tesis de que el significado, la intensión o el sentido determinan la referencia o la extensión de las expresiones. Ante la disyuntiva de tener que abandonar alguna de estas dos tesis, puesto que ninguna noción de significado satisface simultáneamente estos dos requisitos, Putnam opta por rechazar que los significados sean entidades mentales, y deja intacta la herencia analítica de que los significados determinan la referencia. En su caso, el significado determina la extensión, como él mismo señala, “por construcción”, al ser la extensión un componente más del significado, junto con los estereotipos y los marcadores sintácticos y semánticos. Por lo tanto, estamos ante una nueva concepción del significado, que, por un lado, introduce como uno de sus elementos la propia extensión de las expresiones, y, por otro lado, introduce en la determinación de la referencia elementos sociales, históricos y pragmáticos. La determinación del referente se produce entonces por medio de una cadena comunicativa que, como señaló Kripke para los nombres propios, se remonta al momento de la imposición del nombre; pero también, como indicó Putnam para los nombres de géneros naturales o artificiales, tiene lugar por medio de división del trabajo lingüístico, donde juega un papel fundamental la decisión de los expertos. Todos estos aspectos nuevos en el tratamiento del significado ponen de relieve que el significado es un producto social, igual que la determinación de la referencia. Según Putnam o Burge, ya no es necesario que los hablantes tengan una idea absolutamente clara y precisa de lo que las palabras significan. Es suficiente que compartan un estereotipo, que recoja determinados aspectos considerados básicos en la comunidad de hablantes, para que podamos decir de ellos que conocen el significado de las expresiones de su lengua.

Sin embargo la Tradición Analítica, a pesar de seguir en muchos sentidos las pautas marcadas por la Filosofía Analítica, se interesó por determinados aspectos del lenguaje en conexión con el fenómeno de la intencionalidad, el racionalismo y el funcionamiento del cerebro. Los usos del lenguaje nos permiten construir oraciones en las que expresamos nuestras creencias, nuestros deseos, nuestras emociones. Estas oraciones son quizás las que más dificultades presentan a la hora de estudiarlas y comprenderlas, especialmente porque el principio de sustituibilidad *salva veritate* no se mantiene en estos contextos, y como buenos herederos de la Filosofía Analítica, los filósofos de la Tradición Analítica se sienten en la obligación de dar una explicación a este hecho. Pero no sólo la necesidad de dar cuenta de este tipo de fenómenos lingüísticos tuvo que ver en el interés que despertó esta perspectiva más interna de abordar el lenguaje. La revolución lingüística promovida por Chomsky, pareja a su renovación del cartesianismo y del

racionalismo, favoreció la investigación de los aspectos mentales y biológicos vinculados con el pensamiento y el lenguaje. El lenguaje ya no es tan sólo un hecho social, sino que también es -y para Chomsky, sobre todo- un órgano mental que permite expresar pensamientos. La lingüística es, entonces, una rama de la psicología cognitiva. Por otro lado, la creatividad lingüística, destacada especialmente por Chomsky, está directamente vinculada a la recursividad característica de los lenguajes naturales, y es el objeto de estudio propio de la gramática, y no de la actuación lingüística. Desde este punto de vista, pues, no es de extrañar la enorme importancia que Chomsky le dio a la sintaxis de los lenguajes naturales, y la defensa de su autonomía frente a la semántica. Aunque lejanamente en el tiempo y en la tradición, este interés de Chomsky viene a coincidir con el interés que los filósofos analíticos mostraron por la sintaxis y la forma lógica del lenguaje.

Las estrechas relaciones que Chomsky estableció entre lingüística y psicología, y especialmente la hipótesis de Fodor acerca del lenguaje del pensamiento, entre otras, están propiciando un interés cada vez mayor por los aspectos mentales vinculados a la comprensión y expresión de pensamientos a través del lenguaje. La filosofía de la mente y las ciencias cognitivas están ocupando un lugar de honor en el conjunto de los saberes. Como indica Burge, no sólo muchos filósofos son de la opinión de que la filosofía del lenguaje parece que ya realizó su trabajo, sino que también la mayor parte de los problemas filosóficos tratados por ella acaban apuntando a la filosofía de la mente.

La Filosofía Analítica abordó las implicaciones filosóficas del lenguaje desde una perspectiva externalista en las dos orientaciones que la caracterizan: la lógico-semántica y la pragmática. En ambos casos sus estudios obtuvieron éxitos considerables, proporcionando una mejor comprensión tanto de los lenguajes naturales como de los lenguajes formales. El análisis sintáctico resultó a todas luces el más accesible y satisfactorio en sus resultados. Las cuestiones semánticas mostraron una complejidad mucho mayor, y una dificultad enorme a la hora de establecer límites, de conseguir definiciones precisas y de proporcionar sistematizaciones. Una clara prueba de ello la tenemos en que todavía no contamos con una definición unánime de lo que es el significado. Con la incorporación de la dimensión pragmática se completó el otro aspecto relevante en el estudio del lenguaje y del significado. La Filosofía Analítica prestó atención a todas estas perspectivas, aunque de un modo aislado y sin relación entre ellas, esto es, o se concentraba en la dimensión sintáctica, y dejaba a un lado la semántica y la pragmática; o se concentraba en la semántica, obviando los elementos pragmáticos. Así aunque el lenguaje se investigaba

desde todas estas perspectivas no se establecieron las relaciones pertinentes entre ellas. La Tradición Analítica es más promiscua en este sentido, y tiene en cuenta de manera más global e interactiva todos estos aspectos. Kaplan y Perry, por ejemplo, estudiaron los demostrativos, expresiones típicas de la interrelación entre semántica y pragmática, sin por ello dejar de recurrir al aparato formal, combinación impensable en la dimensión pragmática de la Filosofía Analítica. Ahora se tiene la impresión de que el estudio del lenguaje, y del significado, queda completamente abordado cuando el desarrollo posterior de la Filosofía Analítica, esto es, la Tradición Analítica tiene en cuenta todas las dimensiones interesantes del mismo, esto es, la sintáctica, la semántica y la pragmática. Sin embargo queda aún pendiente el gran reto de la comprensión del significado desde la base materialista y mentalista. Es aquí donde la Filosofía de la Mente aporta nueva savia a la Tradición Analítica.

REFERENCIAS

- Acero, Juan José: *Lenguaje y filosofía*, Ed. Octaedro, Barcelona, 1993.
Burge, Tyler: "Philosophy of Language and Mind: 1950-1990", *The Philosophical Review*, Vol. 101, nº 1, 1992, pp. 3-51.
Hierro S. Pescador, José: *Significado y verdad*, Ed. Alianza, Madrid, 1990.
Wittgenstein, Ludwig: *Tractatus Logico-Philosophicus*, Ed. Alianza, Madrid, 1973.
Wittgenstein, Ludwig: *Investigaciones filosóficas*, Ed. Crítica, Barcelona, 1988.

M^a Uxía Rivas Monroy
Universidade de Santiago de Compostela
e-mail: lfuxia@usc.es

EL PAPEL DEL LECTOR EN SEMIÓTICA

DIEGO ROSALES GALIÑANES
Instituto Teológico Compostelano

ABSTRACT: The present contribution revises the role of the reader in semiotics of reception. We departure from the model-work theory developed by Umberto Eco in his book *Lector in fabula* and *Interpretation and overinterpretation*. This is a question about the competence of the reader with respect to the decodification of a given text.

The analysis of the role of the reader puts forward some interesting questions from the philosophical point of view. This approach to the role of the reader can be made in two ways: First, by taken the reader as the final receptor of a given information contained in the text of the message. The second way completes the previous one raising the question for the fiability of the reading done, i.e., the question for the overinterpretations.

KEYWORDS: Interpretation, overinterpretation, intentio auctoris, intentio operis.

1.- Introducción

La finalidad del presente trabajo es propedeútica y consiste en resaltar la importancia de la metodología en filosofía. Por metodología no entiendo la reflexión sobre los métodos generales *-modi sciendi-* ni la reflexión sobre las maneras particulares de conducirse en las disciplinas particulares sino el desarrollo de procedimientos concretos. En este punto se debe hacer el inciso de que el término “procedimiento” se utiliza en el sentido técnico dado en los modernos estudios de calidad empresarial.

La novedad de esos procedimientos es que deben cumplir con su misión heurística tradicional -su orientación a descubrir la verdad de las cosas- respecto al objeto de estudio y, a mayores, tendrán en su salida didáctica una prueba de validez. De otra forma: un procedimiento metodológico debe cumplir como técnica con las condiciones clásicas de adaptarse a las condiciones noéticas del sujeto y a la características ontológicas del objeto, pero, además, debe estar especificado de forma tal que sea posible transmitirlo de forma análoga a cómo se transmitía el conocimiento técnico en la filosofía aristotélica.

Estas matizaciones no son algo baladí puesto que lo que aquí se pone en juego es el estatuto de la filosofía -analítica o de cualquier otro tipo- no como elemento del conjunto de saberes generales sino como oficio. El punto

donde deben caer los esfuerzos de mayor peso es en el desarrollo de procedimientos didácticos que permitan la transmisión a terceros de los procedimientos heurísticos.

R. Nozick apuntó el problema anterior en su libro **La naturaleza de la racionalidad** cuando dijo: *“Resulta un poco sorprendente que, al entrenar a los estudiantes de filosofía, los profesores no hagan ningún esfuerzo serio para formular reglas tentativas de este tipo para conducir la empresa intelectual. Los estudiantes solo reciben productos acabados -libros y artículos- y el ejemplo de sus profesores como analistas y comentadores de los mismos, para ser después abandonados a representarse por sí mismos el modo en el que pudieron ser realizados”* El texto de Nozick permite traer a la palestra un tema urgente para cualquier docente y sugerente para cualquier profesional de la filosofía. De este texto podemos obtener las siguientes pistas:

1.- La filosofía -en su parte docente- consiste en la lectura y comentario de textos. El modelo de análisis le viene dado al alumno por el ejemplo del profesor y por el principio de confianza que le indica al primero que el último constituye un patrón adecuado. El modelo de lectura se deriva de modo incierto del modelo de análisis hasta el punto en que, según Nozick, el alumno queda relegado a una tarea con un gran componente de imaginación inventiva.

2.- El problema se agrava en el momento en el que el alumno se encuentra ante el problema de la producción. El alumno no sabe el cómo de la producción de los textos con los que debe trabajar y con esto pierde la posibilidad de realizar textos propios en el futuro por la técnica más simple que es la de copia.

Un caso concreto que nos permitirá ver el problema de la desconexión de los métodos heurísticos de los métodos didácticos y de la necesidad de desarrollo de procedimientos operativos es el problema de la lectura. Este problema es relevante porque:

1.- El texto en cualquiera de sus formas es la piedra de toque indispensable para llevar a término la mayor parte de las tareas asignadas tradicionalmente a la filosofía. Dada su importancia parece extraña la ausencia de capacidades adecuadas en las personas que manipulan este material para que esta labor sea “correcta” Esto es, la falta de una metodología que permita la separación entre una labor bien hecha y otra que no está bien hecha.

2.- Jamás ha existido una cantidad de publicaciones de “segundo orden” mayor que en la época actual. Por publicaciones de segundo orden entiendo todos aquellos trabajos que son esencialmente comentarios de otras obras. El problema de estas lecturas de segundo orden es que sucede que en ocasiones dos trabajos presentan lecturas incompatibles de una misma obra. Conectando con el punto anterior parece de sentido común admitir en principio que una de las lecturas se acerca al sentido del texto más que la otra.

De esta forma se puede hablar de interpretación/traducción adecuada para una obra que se acerca al sentido original del texto y de sobreinterpretación a una lectura de está alejada en exceso de ese sentido original del texto.

Abordaremos en lo sucesivo el problema de la lectura dentro del ámbito de la semiótica haciendo nuestros los conceptos de expuestos por U. Eco en su obra **Lector in fabula** y recurriendo a la definición etimológica del término “interpretación” que la sitúa como correlato de los términos “mediación”, “traducción” e incluso “adivinación”.

2.- El lector en semiótica

Elegimos abordar el problema de la lectura desde la semiótica porque es una disciplina interdisciplinar. Esta naturaleza “cubista” de la disciplina resulta ventajosa ya que de forma análoga al movimiento pictórico aludido permite la coincidencia de distintos puntos de vista en un sólo espacio, lo cual genera una enriquecedora cantidad de puntos de vista. La semiótica es, por otra parte, una disciplina joven -con mucho éxito popular en momentos puntuales- que todavía está construyendo su propio lenguaje y esto la hace permeable en su dominio y susceptible de innovación dentro del mismo.

Dentro del dominio de la semiótica entran campos en común con la filosofía del lenguaje y con la tradición analítica de la filosofía. En el nacimiento de esta disciplina están personajes como Peirce, Morris, etc. que resultan elementales en una exposición completa de esta tradición. Pero además de compartir autores la semiótica también tiene un área común con la tradición analítica porque hace suyos todos los problemas semánticos como la polisemia, la connotación, la determinación del sentido y el referente, etc. Si como afirma U. Eco en su libro **Tratado de semiótica general** la semiótica se ocupa de todo aquello que sirve para mentir el problema del significado por limitación y el de la verdad y la falsedad por extensión se vuelven centrales en esta disciplina.

Quizás la singularidad más evidente de la semiótica frente a la filosofía analítica es que ésta última se centra en los códigos, mientras que la semiótica trabaja también el campo de la comunicación. Es dentro de este último campo donde es adecuado plantear el problema de la lectura puesto que el proceso de lectura no deja de ser un proceso de comunicación al uso -un proceso de comunicación en el que la fuente de datos coincide con el emisor, por otra parte. El interés del presente trabajo consiste en el estudio de la decodificación del mensaje por parte del receptor, el establecimiento de criterios para la evaluación de esa decodificación y la posibilidad de describir tanto el proceso como los criterios en forma de procedimientos que puedan plasmarse en “recetas” didácticas. En este punto es claro que el estudio propuesto entra dentro de una semiótica de la recepción.

2.1.- Descripción del proceso lector

La lectura está formada por el conjunto de acciones que se realizan al leer; y leer apunta etimológicamente al término griego “lego” que significa “escoger” y al verbo latino “legere” que apunta en su origen a vocablos como “recorrer”, “apropiarse”, “entresacar”, etc. La lectura es un proceso complejo de múltiples dimensiones -físico, psicológico, gnoseológico y social- en el cual un receptor recorre con la vista una serie de signos para apropiarse de los sentidos de dichos signos para entresacar las informaciones que estos sentidos contienen. En este proceso el receptor o lector no es pasivo respecto a las informaciones que le proporcionan esos sentidos de los signos sino que construye nuevos sentidos de los signos a partir de los sentidos primeros.

Por esto, el proceso lector está formado por el objeto de lectura -el escrito-, la percepción de dicho objeto y el proceso psicológico que ocurre en el lector al asimilar el sentido del texto, evaluarlo y completarlo. La buena marcha de este proceso depende de las siguientes condiciones:

- La falta de dificultades en la percepción del objeto.
- El conocimiento del código de signos por parte del receptor.
- El adecuamiento del sentido dado por el receptor respecto del objeto.

El punto más relevante desde la óptica semiótica es el tercero. Cualquier acto de lectura de un hablante con la capacidad de leer respecto a un escrito en su lengua natural presupone la competencia de poder identificar los signos del código de esa misma lengua natural. Sin embargo, los problemas acerca de la adecuación del sentido son ciertamente más complejos puesto que el otro polo de la relación de adecuación antes de ser claro es más bien confuso. De otra forma: el sentido obtenido por el lector debe adecuarse al sentido que el autor del escrito ha puesto en el mismo, al sentido propio de la obra sin entrar en la consideración del autor, a la relación de dos sentidos obtenidos por el receptor sin más, la falta total de cualquier relación de adecuación, etc.

2.2.- Los paradigmas de la interpretación

Desde un punto de vista clásico el proceso de lectura en filosofía equivale a la interpretación del texto. Interpretar es realizar una mediación entre el sentido expreso del texto y el sentido latente en el mismo. Esta doble perspectiva entre el sentido expreso y el sentido latente del texto es lo que va a elaborar la moderna teoría de la interpretación que nace con Schleiermacher en el siglo XIX.

La dificultad para realizar la definición de ese sentido latente va a traer consigo el fenómeno de la sobreinterpretación. La sobreinterpretación va a estar mas o menos presente en función del paradigma utilizado para la exégesis del texto. A grandes rasgos podemos distinguir tres grandes paradigmas en cuanto a la interpretación de los textos se refiere según se alejen de la intención pretextual -esto es, la intención del autor- o

permanezcan cercanos a ella. Según esto podemos distinguir el paradigma “detectivesco”, el paradigma objetivista y el paradigma reconstruccionista.

El paradigma “detectivesco” es el más útil e intuitivo a la hora de dar a entender la labor que debe desempeñar la interpretación a un alumno. De igual forma a como un detective realiza una pesquisa para encontrar pistas para resolver el caso criminal que le ocupe, un investigador -el que realiza una investigación, del latín *in-vestigium*, buscar pistas- debe encontrar los vestigios que le permitan reconstruir el sentido dado por el autor a su texto. La gran ventaja de esta idea reguladora es que la investigación se presenta al alumno como un problema de dirección única: existe un sentido originario -el pensamiento expresado por el autor en el texto- y la tarea del intérprete es recuperarlo para su público liberándolo de toda mistificación, error o adulterio. El paradigma se propone sobre la propia idea de método -camino (*odos*) hacia (*meta*)- y esto le dota de simplicidad explicativa.

Las desventajas de este paradigma también resultan evidentes. Este paradigma supone el postulado de un sentido único en el texto que no es ni mucho menos claro. Antes bien, puede atribuirse a un convencionalismo del intérprete que resulta perjudicial en el sentido de que elimina más interpretaciones empobreciendo así las posibilidades para sugerir del texto. Este paradigma se construye sobre el mito de la objetividad -existe lo que buscamos, es real de la misma manera en la que son reales los objetos del mundo físico- y tan solo hemos de guiarnos por el mapa para no perdernos en el territorio en el que nos encontremos.

Además, el autor del texto puede no ser consciente de los sentidos encontrados por el intérprete y sin embargo, al conocerlos, aceptarlos como plausibles. Más aún, el autor rara vez dota a su texto de un sentido tan acabado que excluya posibles sentidos alternativos. En un caso límite como reconoce Eco [Eco97, 80 y ss.], la labor del autor se restringiría a la negación de posibles sentidos, a descartar interpretaciones. Ahora bien, la intención del autor -*intentio auctoris*- no es canon de medida para dar por adecuadas unas interpretaciones y por inadecuadas otras. Eco entiende que esa misma interpretación del autor puede ser equívoca o irrelevante para determinar la intención del texto.

Esta irrelevancia del papel del autor -el emisor latente tras el objeto de lectura- lleva a Eco a postular a la propia propia obra como canon de medida. La obra de arte es abierta pero posee su propia *intentio operis* -su sentido propio- que la hace independiente del autor pero que la cierra lo suficiente para no permitir cualquier lectura de ella. Esto se completa con la postulación por parte de Eco de su noción de lector modelo. Una obra cualquiera es leída por un lector competente pero de una forma que no puede ser determinada como la adecuada. El lector modelo es aquel que alcanza el sentido que la obra tiene como originario, conecta con la *intentio operis*.

Este paradigma de Eco posee la ventaja de que se hace independiente de un lector y una obra concretas y con esto gana en objetividad respecto del paradigma “detectivesco”. Las nociones de obra modelo y lector modelo parecen seres del tercer mundo postulado por K. Popper en su obra **Conocimiento objetivo**. Es un paradigma más rico conceptualmente pero menos intuitivo y claro. De igual forma que el anterior este paradigma se sustenta sobre el mito de la objetividad al suponer un sentido único válido intersubjetivamente.

El tercer paradigma representa el máximo alejamiento respecto a la intención auctoral. El texto es independiente de toda intención pretextual y, a la vez, es capaz de generar sentidos por sí mismo. Dentro de este paradigma aparecen dos postulados:

- No hay un sentido del texto único en el texto. El sentido del texto cambia y es inestable en el sentido en el que habla J. Derrida. Esto supone un alejamiento de la postura de Eco ya que el texto no podría desde esa inestabilidad de sentido generar una lectura (obra/lector) modelo.

- El texto es en potencia productor de infinitos sentidos. Un texto con capacidad para generar un mayor número de lecturas se supone más rico que otro texto que no tenga esa cualidad. Sin embargo esa capacidad de “semiosis ilimitada” trae consigo el problema del límite, esto es, en que momento las lecturas ya están tan alejadas de todo sentido propio del texto que se constituyen en lecturas inadmisibles o sobreinterpretaciones.

Este tercer paradigma tiene sus representantes en los deconstruccionistas y en la Nueva Crítica norteamericana -J. Crowe, A. Tate, etc. Estas corrientes de críticos literarios parten de la autonomía del texto caracterizándolo como algo autónomo y de sentido autosuficiente. En base a esto se elaboran teorías coherentes con los supuestos sentidos latentes del texto y la propia coherencia de las exposiciones justifica las mismas. La intención pretextual del autor pasa a tener grado de falacia y denominarse “falacia intencional”. Esta falacia y todo criterio que limite nuevas interpretaciones a ser acusado de tener un enfoque políticamente conservador y de naturaleza autoritaria.

Las ventajas de este paradigma radican en que provoca la rápida sucesión de lecturas y admite la contradicción entre ellas. Debido al número de lecturas que contempla podemos entender que es un paradigma fértil y democrático. Además, vence el mito de la objetividad que afectaba a los dos anteriores -sobre todo al paradigma “detectivesco”- que suponía la existencia de una lectura/interpretación correcta que podría ser alcanzada en primer lugar, para después ser utilizada como criterio de demarcación ante otras lecturas. Un texto absolutamente abierto es una fuente de sugerencias y preguntas infinitas, pero aquí está precisamente su talón de Aquiles puesto que en esa infinitud de lecturas se vuelve el texto difícilmente manejable.

La principal desventaja que aqueja a este tercer paradigma es que si bien es muy fecundo como fuente de lecturas y de teorías que puedan dar coherencia a las mismas se presenta como algo inmanejable desde el punto de vista didáctico. Este tercer paradigma parece ejemplificar el “dadaísmo epistemológico” expresado por el lema del filósofo de la ciencia P. K. Feyerabend en su famoso lema “todo vale”. Dentro de este tercer paradigma ocurren dos cosas: una, no existe la posibilidad de elaborar métodos didácticos puesto que el mero intento de hacerlo se vuelve absurdo. Si la norma es que no hay normas, ¿qué sentido tiene intentar explicitar una? Dos, el problema de R. Nozick pierde todo sentido puesto que los alumnos no están abandonados a su suerte y al ejemplo como comentadores de sus profesores: basta con que expresen sus opiniones para que se muestren tan capaces de producir un discurso tan válido como el de sus formadores. Esto se muestra excesivo a la intuición.

3.- La sobreinterpretación

Una sobreinterpretación es una lectura de un texto que se muestra inadmisibles para este. De esta forma podemos entender que una sobreinterpretación es una relación entre el texto y el lector en el que la relación mediadora -la lectura- ha tomado importancia por sí misma. En la sobreinterpretación se produce una traducción del texto original que traiciona el sentido del mismo.

Un posible origen de las sobreinterpretaciones puede encontrarse en la extrema actitud de sospecha de algunos críticos respecto a los textos. Estos críticos son tildados por Eco [Eco97, 40-55] como “los adeptos al velo” - en el sentido de la verdad como *aletheia*- en tanto que siempre creen que la verdad está por descubrir, que siempre hay un velo que descubrir para ver la verdad bien redonda que nos muestra la diosa de forma análoga al poema de Parménides. Esta actitud de búsqueda de sentidos ocultos y herméticos, de deshacer los secretos taumatúrgicos que contiene de forma latente -y sin final, podríamos añadir- el texto provoca la aparición de las interpretaciones exageradas.

Un ejemplo sobre interpretación es el que cita Eco sobre la lectura de Dante por parte de Rosseti. Eco escribe en [Eco97, 68-70]: “*Rosseti se sorprende de que aparezcan en Dante referencias a la cruz, la rosa y el pelicano. Las razones de la presencia de estas palabras son evidentes. En un poema que habla de los misterios de la religión no es sorprendente que tarde o temprano aparezca el símbolo de la Pasión. Sobre la base de una antigua tradición simbólica, el pelicano se convirtió muy pronto en el símbolo de Cristo en el seno del cristianismo (y la poesía religiosa y los bestiarios medievales están llenos de referencias a ese símbolo)...Sin embargo, Rossetti también quiere el pelicano, lo encuentra por su cuenta en Paraíso XXV (su única aparición en el poema), claramente en conexión con la cruz, porque el*

pelicano es el símbolo del sacrificio. Por desgracia, no hay ninguna rosa. De modo que Rossetti parte en busca de otros pelicanos. Encuentra otro pelicano en Cecco d'Ascoli (otro autor en torno a quien los adeptos del velo se han devanado los sesos por la sencilla razón de que el texto de L'Acerba es intencionadamente oscuro), y el pelicano de Cecco aparece en el contexto usual de la Pasión. Además, un pelicano en Cecco no es un pelicano en Dante, aun cuando Rossetti intente difuminar esa diferencia menor confundiendo las notas al pie de página. Rossetti cree que ha encontrado otro pelicano en el principio de Paraíso XXIII, cuando leemos sobre el ave que, esperando con impaciencia el amanecer, está posada vigilante entre la espesura de una rama frondosa atenta a la salida del sol para partir en busca de comida para sus crías. Ahora bien, esta ave, en verdad delicada, busca comida precisamente porque no es un pelicano, de otro modo no necesitaría salir a cazar, podría alimentar a su prole con carne sacada de su pecho. En segundo lugar, aparece como un símil de Beatriz y habría sido un suicidio poético que Dante representara a su amada con las extrañas características de un picudo pelicano. Rossetti, en su desesperada y bastante patética partida de caza, encontró en el divino poema siete aves y once pájaros y los adscribe todos a la familia del pelicano; pero los encuentra siempre lejos de la rosa"

Utilizando una metáfora espacial podemos entender que las interpretaciones adecuadas se encontrarán más cerca del sentido original del texto que las sobreinterpretaciones. Continuando con esta metáfora espacial entendemos que cuanto más alejada esté la sobreinterpretación más extraña resultará al propio texto y más fácil será de identificar sin la necesidad de recurrir a conocimientos específicos. Ahora bien, las sobreinterpretaciones problemáticas son aquellas que se "tocan" con las interpretaciones legítimas. Se hace necesario entonces un criterio que permita separar unas de otras. Es obvio que postular un criterio -sea el que sea- es aceptar el marco conceptual de uno de los dos primeros paradigmas expuestos en el apartado anterior -en el caso de Eco esto es claro ya que se va a mover en el marco conceptual del segundo paradigma, de su propia elaboración.

3.1.- Criterios de evaluación

En su libro **Los límites de la interpretación** [Eco98, 15-19] Eco nos proporciona una serie de criterios para evitar todo "*derroche de energías hermeneúticas*" Estos criterios son importantes porque en la opinión del autor no sólo son los mejores posibles sino que tienen ese estatuto porque también son los únicos posibles. Estos criterios son:

- El sentido común.
- El principio de mínimo esfuerzo.
- El consenso de los especialistas.

Estos tres criterios tienen como base la economía. En el artículo “Interpretación e historia” Eco completará la naturaleza crematística de esos principios acudiendo al principio popperiano de falsabilidad. De forma análoga a una teoría científica, las interpretaciones se abandonan por otras con mayor poder explicativo y menos complejidad conceptual; esto es, interpretaciones prototípicas del principio del mínimo esfuerzo. Pero lo cierto es que en el caso de la interpretación también funciona el sentido común puesto que para las interpretaciones “espacialmente” separadas de la intención pretextual no hace falta buscar un experimento crucial que la refute, por seguir con las analogías a la filosofía de la ciencia de Popper.

El consenso de los especialistas actúa en muchos de los casos como el propio autor/emisor del mensaje puesto que establecerán el contexto y las condiciones que proporcionarán validez a una interpretación particular. No obstante, ni este *consensium gentium* tan determinante en la ciencia de nuestros días -pensemos de forma ilustrativa en el papel que H. Putnam achaca a los especialistas que limitan los estereotipos en su artículo “*El significado de significado*”- ni los criterios anteriores son lo bastante determinantes para establecer la partición lógica necesaria que separe con nitidez las interpretaciones de las sobreinterpretaciones.

Estos tres son criterios blandos no son suficientes puesto que da la sensación de que no hay una intersección vacía entre las interpretaciones y las sobreinterpretaciones. El criterio del sentido común parece provenir de una capacidad poco extendida y de complicada transmisión a terceros. El principio del mínimo esfuerzo es un ideal a alcanzar pero rara vez un principio operativo desde el comienzo de una investigación. Y el consenso de los especialistas sólo asegura el aplauso de una comunidad científica particular no la verdad de la interpretación, aunque ambas las consideremos como unidas. A mayores se hace evidente que estos criterios pueden ser expuestos como ideas reguladoras de lo que tiene/debe ser una labor científica pero no pueden ser explicitados en forma de procedimientos que puedan ser enseñados en el aula.

4.- Procedimientos didácticos

La diferencia fundamental entre ciencia y tecnología es que la segunda es susceptible de normalización conceptual (ver [Mandado92]). La normalización tiene dos ventajas *a priori* para cualquier disciplina: una, supone una considerable reducción de la inflación terminológica de la que suelen adolecer las disciplinas; y dos, permite el establecimiento de procesos estandar que pueden ser comunicados con relativa facilidad. En el caso de la filosofía existe una tercera ventaja y es que de las dos anteriores se deriva la posibilidad de evitar lo que R. Rorty denomina la “aspiración fundacionalista” [Eco97, 104-126] o la búsqueda de los fundamentos que permitan la explicación del “ser” -propiedades intrínsecas- de la “realidad” Rorty, como

filósofo pragmatista que es, no distingue entre usar e interpretar un texto. Un texto es un instrumento de cara a la finalidad práctica que se le encomiende y su interpretación sólo tiene sentido en tanto nos permita alcanzar esa finalidad. Para el filósofo norteamericano no tiene utilidad alguna realizar una lectura para desentrañar los vericuetos propios del vocabulario de un texto.

La postura de Rorty es excesiva y hace suyo un pragmatismo menos conciliador que el de W. James. Ahora bien, la idea de considerar el conocimiento como instrumentos es útil desde el momento en el que da pie a la pregunta acerca de si las ideas de la filosofía pueden ser “normalizadas” de forma semejante al utillaje técnico. Desde luego, el paso previo antes de realizar ese proceso de normalización le supondrá a la filosofía una necesaria homologación y diferenciación metodológica respecto a otras disciplinas afines a ella. La homologación en tanto que debe establecer los procedimientos que le son propios y diferenciación en tanto que debe identificar las aportaciones dentro de ella de otras disciplinas, i.e., la literatura, y evaluarlas.

La normalización en el sentido de la calidad empresarial es el establecimiento de normas. En sentido amplio representa la unificación de criterios y la posibilidad de utilizar un lenguaje común. Las ventajas de la normalización son evidentes puesto que garantizan:

- La simplificación conceptual.
- La facilidad de comunicación.
- La optimización y racionalización de los procesos científicos.
- La detección de fallos internos y externos en la teorías.
- La posibilidad de verificar las mismas al tener unos criterios unificados.

La normalización asegura el conocimiento del proceso a todo individuo que intervenga en él y le garantiza la posibilidad de participar en el mismo. Esta participación no sólo garantiza un mayor contacto con el proceso de construcción del conocimiento -Popper apunta en **Conocimiento objetivo** esta idea con el símil de la construcción de la casa. Se aprende más construyendo la casa que estudiándola sin más. Pero lo más importante es que al ser explícitos los procedimientos de construcción estos pueden ser enseñados como son enseñadas las técnicas de cualquier otro oficio.

Quizás estas ideas aparezcan como algo alocado dentro del campo de la filosofía pero en realidad esto no es así. Dentro del campo de la filosofía analítica se utiliza un método -homónimo del nombre de la tradición- todos los días pero jamás se ha hecho el esfuerzo de explicitarlo. Hacerlo es hacer ganar en calidad a la filosofía y abrir el camino para que vuelva a recuperar el puesto que merece dentro de las disciplinas del saber.

BIBLIOGRAFÍA

- Eco, U: *Interpretación y sobre interpretación*. Cambridge University Press, Cambridge, Madrid, 1997.
- Eco, U: *Los límites de la interpretación*. Lumen, Barcelona, 1998.
- Eco, U: *Lector in fabula*. Lumen, Barcelona, 1999.
- Mandado Pérez, E.: *Galicia ante el reto tecnológico*. Lección inaugural de la Universidad de Vigo, 1992.
- Nozick, R.: *La naturaleza de la racionalidad*. Paidós, Barcelona, 1995.
- Popper, K.: *Conocimiento objetivo*. Tecnos, Madrid, 1974.

Diego Rosales Galiñanes
Instituto Teológico Compostelano
e-mail: drosales@fynet.net





DAVIDSON Y TARSKI, ¿PODEMOS REDUCIR SIGNIFICADO A VERDAD?

JORDI VALOR ABAD
Universidad de Valencia

ABSTRACT: My concern in this paper focuses on Davidson's theory of truth for natural Languages. After presenting in a rough outline his reduction of the notion of meaning to that of truth within a tarskian frame, I discuss several difficulties arising from the resulting theory. In particular, I bring out the relevance of Tarski's objections against the possibility of such a project and I draw the reader's attention to the problems that demonstratives and semantical paradoxes pose for Davidson's goals.

KEYWORDS: Meaning, truth, convention (T), indexicals, paradox

Cualquier lenguaje natural es suficientemente rico en recursos como para permitir la construcción de un número infinito de expresiones complejas a partir de un número limitado de elementos básicos y reglas de composición. Por esta razón, uno de los primeros problemas que deberá afrontar una teoría del significado aplicable a un lenguaje natural es el de distinguir, de entre las múltiples expresiones que comprende el lenguaje estudiado, aquellas que son significativas de aquellas que no lo son. El hecho de que el número de expresiones que puede generar un lenguaje natural sea infinito dificulta, sin duda, esta tarea, sin embargo, la existencia de reglas que determinan la formación de la mayoría de expresiones válidas del lenguaje, garantiza que nuestros esfuerzos tendrán algún tipo de recompensa. Con la ayuda de un diccionario y una gramática nos es posible determinar, por ejemplo, cuestiones como el vocabulario de que dispone el lenguaje o las estructuras sintácticas que determinan las oraciones válidas del mismo. Así pues, a partir de un número finito de palabras y de reglas de formación de oraciones, podemos determinar una buena cantidad de expresiones significativas pertenecientes al lenguaje que es objeto de estudio. De alguna manera, y a la vista de este hecho, cabría afirmar con Davidson que "el trabajo de una sintaxis modesta es el de

caracterizar la *significancia* (u oracionalidad)”¹ de las expresiones del lenguaje, pero, por desgracia, y, pese a la gran ayuda que la sintaxis nos brinda en este sentido, nos encontraremos en un sinfín de ocasiones frente a situaciones en las que una oración gramaticalmente correcta no consigue tener significado o incluso frente a situaciones en las que una expresión de dudosa corrección sintáctica resulta, sin embargo, perfectamente significativa².

Con todo, la existencia de un conjunto finito de reglas de composición y elementos lingüísticos básicos es fundamental si queremos tener criterios de decisión respecto a la validez y significancia de ciertas expresiones. Que a veces las reglas no basten o que existan excepciones a ellas no es motivo suficiente para desecharlas. Además, si no dispusiéramos de una gramática finita, nos resultaría prácticamente imposible determinar los límites de nuestro lenguaje o incluso llegar a aprenderlo alguna vez. Por todas estas razones podemos concluir que la gramática y el vocabulario de una lengua representan una ayuda inestimable y eficaz a la hora de determinar la significancia de las expresiones de un lenguaje natural, si bien es cierto que no basta para determinar completamente (si es que ello es posible) el conjunto de todas sus expresiones significativas.

Pero, si el problema de determinar las expresiones significativas de un lenguaje es espinoso, el problema de determinar el *significado* de todas estas expresiones lo es aún más, y es precisamente ésta la preocupación central de toda teoría semántica que se precie y el asunto que nos ocupará en lo que resta de escrito.

Por lo dicho hasta ahora, uno de los requisitos que presumiblemente debería satisfacer una teoría semántica es el de explicar cómo el significado de una expresión compleja depende de su estructura o, en otras palabras, cómo contribuyen al significado general de la expresión sus partes constituyentes. Una teoría que no satisfaga este requisito sería incapaz de explicar similitudes de significado entre oraciones distintas a partir de estructuras compartidas y, sobre todo, ignoraría el hecho de que los hablantes de una lengua *construyen* oraciones sobre la base de expresiones más simples y que, al hacerlo, siguen no sólo reglas sintácticas sino también reglas semánticas que determinan la aportación de las expresiones empleadas al significado general de la oración.

¹ “Verdad y significado” (en Davidson, 1984, pág. 21)

² Un ejemplo de oración correcta aunque no significativa podría ser “El pan subió a la estrofa en dos países” cuya estructura es idéntica a la de la oración perfectamente significativa “El gato subió a la mesa en dos segundos”. Un ejemplo, en cambio, de expresión perfectamente significativa pero de corrección gramatical más bien incierta sería por ejemplo el verso más famoso de uno de los sonetos más célebres de Quevedo, aquel que dice: “soy un fue, y un será, y un es cansado” donde el artículo indeterminado y el adjetivo conciertan con formas conjugadas de verbos.

Por otra parte, si no fuésemos capaces de proporcionar una explicación estructural del significado nos encontraríamos en una situación similar a la descrita anteriormente, cuando planteábamos la posibilidad de un lenguaje natural sin una gramática finita. En esas circunstancias, nos resultaría del todo imposible aprender un lenguaje cuyos significados “básicos” son infinitos y no pueden ser descritos estructuralmente.

En el presente trabajo, pretendo discutir la validez de la teoría semántica que, con el compromiso de satisfacer el requisito ya mencionado, Davidson presenta en “Verdad y significado”. A tal efecto presentaré brevemente en la siguiente sección la teoría de Davidson y discutiré en las posteriores algunos aspectos de la misma que no me resultan convincentes.

1. Solución de Davidson a los problemas de una teoría semántica

Si para cada palabra del lenguaje, o al menos para cada expresión significativa “simple”, pudiéramos encontrar un “significado”, una entidad asociable a dicha expresión de forma inequívoca, entonces el problema de la teoría que buscamos estaría resuelto. A través de un diccionario de “significados atómicos” y una gramática que regule su composición resultaría factible elaborar una teoría del significado de carácter estructural. Desafortunadamente la empresa no va a ser tan sencilla. Dejando a parte la cuestión de la polisemia (habitual en muchas palabras del lenguaje), una de las primeras dificultades con las que tropezaría una teoría de este tipo es la de determinar el papel que juegan las palabras en contextos donde no parece ser su “significado” usual (entendido aquí como referencia) el que contribuye al significado general de la expresión en la que aparecen. Éste sería el caso, por ejemplo, de las oraciones de creencia o de algunas oraciones en estilo indirecto³. Otra dificultad de índole distinta es la que supone la paradoja del análisis. Si realmente podemos analizar el significado de una oración en una serie de significados atómicos en los que se descompone, unos correspondientes a objetos y otros a relaciones que los unen, entonces cabrá preguntarse cómo se relacionan o concatenan entre sí dichos átomos de significado para formar complejos y ello nos llevará a la postulación de nuevas entidades (relaciones) que expliquen este hecho. La consecuencia inmediata es un regreso al infinito.

En sus artículos “Sobre sentido y referencia” y “Sobre concepto y objeto” Frege ofreció una solución brillante a ambos problemas. Por un lado, asoció con cada expresión significativa dos entidades: su referencia y su

³ Ciertos enunciados de identidad como “El lucero de la tarde es el lucero de la mañana” presentarían problemas similares.

sentido con lo cual podía explicar cómo en determinados contextos era el sentido de una expresión, y no su referencia habitual, el que contribuía al significado general del complejo. Por otro lado, construyó una ontología con dos tipos de entidades: objetos y conceptos, donde los segundos, a diferencia de los primeros, eran entidades esencialmente incompletas o insaturadas que no debían ser referidas a través de sustantivos o expresiones apropiadas para designar objetos. De esta manera podía salvar la paradoja del análisis (que descansa en la cosificación de las relaciones) aunque a un precio muy alto, el de no poder hablar de los conceptos, lo cual no deja de ser paradójico. La teoría del significado de Frege, también de carácter atomista, resuelve algunos de los interrogantes planteados por la teoría inicial pero, como hemos visto, no está exenta de problemas.

Después de discutir éstas y otras cuestiones Davidson rechaza la viabilidad de una semántica atomista argumentando que no es preciso apelar a entidades tales como “significados” para ofrecer una teoría semántica constructiva que tenga en cuenta la estructura de las expresiones lingüísticas. Así pues no sólo rechaza la existencia de significados particulares asociados a palabras concretas sino también la existencia de significados asociados a oraciones y propone a cambio una perspectiva holista del significado: “Si el significado de las oraciones depende de su estructura y entendemos el significado de cada elemento de la estructura como una abstracción a partir de la totalidad de oraciones en las que figura, entonces tan sólo podemos dar el significado de una oración cualquiera (o una palabra) dando el significado de todas las oraciones y palabras del lenguaje”⁴. Es el lenguaje en su conjunto la unidad mínima de significado.

Pero, ¿cómo podemos articular una teoría semántica que sea holista (prescinda de significados atómicos) y pueda explicar a la vez el significado de las expresiones lingüísticas a partir de su estructura? La respuesta de Davidson consiste en reducir la noción de significado a la de verdad, ofreciendo una teoría de corte tarskiano para esta segunda.

Ya que la noción de significado nos ha conducido a la postulación de entidades problemáticas y a la formulación de las paradojas que hemos tratado con anterioridad, Davidson propone sustituir la expresión “o significa que p” por la expresión:

(V) o es verdadera si y sólo si p.

donde “p” y “o” son, en ambos casos, variables a sustituir por una oración y una descripción estructural de dicha oración respectivamente. Una teoría que cumpla la convención (V) de Tarski, es decir, una teoría capaz de demostrar

⁴ “Verdad y significado”, en Davidson, 1984, pág. 22

para cualquier oración del lenguaje el bicondiciondional (V) correspondiente, es una teoría válida del significado para ese lenguaje en la medida en que “funciona dando las condiciones necesarias y suficientes para la verdad de toda oración y dar las condiciones de verdad es una manera de dar el significado de una oración”⁵.

La teoría tarskiana de la verdad parece satisfacer los requisitos de una teoría a la vez estructural y holista del significado en dos sentidos: a) la determinación de las condiciones veritativas (y por tanto del significado) de una oración compleja descansa en la determinación de las condiciones de verdad de las oraciones que la constituyen. Dado el carácter recursivo de la definición de verdad, podemos asignar valores de verdad a oraciones complejas (oraciones con conectivas lógicas) a partir de las tablas de verdad y de los valores de verdad que hemos asignado a sus componentes; b) La asignación general de valores de verdad de acuerdo con las leyes de la lógica es una tarea que involucra a todas las oraciones de un lenguaje so pena de incurrir en alguna inconsistencia. El bicondicional “‘la nieve es blanca’ es verdadera si y sólo si la nieve es blanca” no da por sí mismo el significado de la oración ‘la nieve es blanca’ más de lo que pueda hacerlo el bicondicional “‘la nieve es blanca’ es verdadera si y sólo si la hierba es verde”. No se trata de poner en el extremo derecho del bicondicional una oración que sea verdadera siempre que lo sea la oración mencionada en el extremo izquierdo, de lo que se trata es de que el conjunto de los bicondicionales de Tarski contruidos sea internamente consistente y es más probable que esto ocurra si utilizamos el primer bicondicional para ‘la nieve es blanca’ que si utilizamos el segundo.

Entre las muchas ventajas que Davidson atribuye a esta teoría se encuentran las de no utilizar términos semánticos que no hayan sido previamente definidos, ya que la noción de verdad se define a partir de la noción semántica de satisfacción, perfectamente definida entre oraciones y secuencias infinitas de objetos, y la de proporcionar, a través de la relación de satisfacción entre oraciones y secuencias, un sentido relevante en el que una teoría de la verdad puede ser una teoría de la correspondencia⁶.

Sin embargo, el propio Davidson reconoce que el intento de transportar la teoría tarskiana de la verdad del ámbito de los lenguajes formalizados al de los lenguajes naturales entraña ciertos problemas. Esencialmente tres:

1) La presencia de elementos *indéxicos* (demostrativos, tiempos verbales, adverbios,...) que hacen variar las condiciones de verdad de una oración de uno a otro contexto de emisión. Davidson propone como solución modificar la

⁵ op. cit. pág. 24.

⁶ “True to the Facts”, en Davidson, 1984, capítulo 3.

convención (V) para relativizar la noción de verdad a la de oración, tiempo de emisión y hablante. Así pues tenemos:

(V') La oración *o* es verdadera si y sólo si es emitida por el hablante *h* en el tiempo *t* y *p*. (Donde "*p*" es una variable a sustituir por una oración cuyas condiciones de verdad son idénticas a las de la emisión de la oración "*o*" por "*h*" en "*t*").

2) Las *paradojas semánticas*. Davidson sólo sugiere que "surgen cuando el rango de los cuantificadores en el lenguaje objeto es demasiado generoso"⁷.

3) La ausencia de una especificación de la forma lógica de ciertas oraciones (imperativas, contrafácticas,...) cuyas condiciones de verdad (si tienen) no están claras, y la necesidad de una descripción sistemática (estructural) de la forma en que contribuyen al significado de una oración adjetivos atributivos, adverbios, términos de masa -agua, fuego,...-, etc.

El tercer apartado, más que un problema, encierra todo un ejército de ellos, ya que es bastante incierto que el significado de todas las estructuras lingüísticas que aparecen en él pueda reducirse por completo a la noción de verdad. Sin embargo, vamos a dejar de lado este apartado y nos vamos a centrar en los otros dos que, al menos, están mejor delimitados. ¿Son las respuestas de Davidson satisfactorias?

2. Los demostrativos y la convención (V)

Un primer inconveniente que presenta la teoría de la verdad de Davidson frente a la de Tarski es que sus bicondicionales no son tan obvios como los de aquella, de hecho, son radicalmente diferentes. Tarski definió la noción semántica de verdad en relación a las *oraciones* de un lenguaje formalizado: en el extremo derecho de los bicondicionales (V) aparece siempre *una oración* y en el izquierdo una descripción estructural de la misma. Davidson, en cambio, no puede utilizar a las oraciones como portadoras de verdad ya que la presencia de elementos indéxicos en los lenguajes naturales determina diferentes condiciones veritativas para una misma oración en contextos distintos. De hecho, Los demostrativos pueden incluso impedir la correcta instanciación de un bicondicional: "'Esta oración es corta' es verdadera si y sólo si esta oración es corta", no es, obviamente, una instancia válida del bicondicional (V) para la oración 'esta oración es corta' porque la referencia del demostrativo ha cambiado contextualmente. El problema no se resuelve utilizando una descripción estructural de la oración alternativa a la que nos ofrece el entrecomillado, sino empleando en el extremo derecho del bicondicional una expresión que tenga *la misma referencia* que el demostrativo

⁷ "Verdad y significado", en Davidson, 1984, pág. 28

de la oración cuyo bicondicional queremos construir. Así pues Davidson se ve forzado a reformular los bicondicionales de Tarski, aunque su formulación alternativa a penas ofrece una solución en el caso que nos ocupa: “‘Esta oración es corta’ es verdadera si y sólo si es emitida por el hablante *h* en el tiempo *t* y *aquella* oración es corta” plantea más dificultades de las que resuelve. En primer lugar, y a diferencia de lo que ocurría en la teoría tarskiana, lo que aparece en el extremo izquierdo del bicondicional *no es una descripción estructural* de lo que ocupa el extremo derecho. A la izquierda del bicondicional se describe estructuralmente una oración, a la derecha, en cambio, se intenta captar una enunciación de la misma. La reformulación de la convención (V) de Tarski se parecería más a la original si en el extremo izquierdo encontrásemos una descripción estructural de lo que aparece en el extremo derecho, pero, por desgracia, esto no es posible porque lo que aparece en el extremo derecho es un enunciado, una entidad cuya determinación depende de factores extralingüísticos y extragramaticales (contextos, tiempos,...). En segundo lugar se nos plantea el problema de cuántas enunciaciones posibles hacen verdadera a una oración y si necesitamos un bicondicional (V’) para cada una de ellas. Y, en tercer y último lugar, surge la cuestión de cómo podemos determinar la contribución que hacen los demostrativos y los indécicos al significado general de las expresiones en las que aparecen. El aspecto más conflictivo de los bicondicionales de Davidson es el lugar que ocuparía la “*p*”, esta variable debe sustituirse por una oración cuyas condiciones de verdad han de reflejar las de la emisión de la oración que origina el bicondicional y esto no se consigue simplemente haciendo referencia a un hablante y un tiempo. Como ha quedado patente en el ejemplo anterior, a veces nos veremos obligados a sustituir un demostrativo por una expresión que contribuya al significado del bicondicional del mismo modo en que lo hace el demostrativo en la oración original. Pero, ¿es esto posible en líneas generales? Supongamos que sí, supongamos que es siempre posible determinar sistemáticamente la contribución que hace un demostrativo al significado de la oración en la que aparece y sustituirlo, consiguientemente, por una expresión equivalente. ¿No equivaldría esto a dar el significado del demostrativo? Si fuese posible determinar la contribución semántica de cada palabra en el contexto en el que aparece (como sugieren los bicondicionales de Davidson), entonces sería posible construir una teoría atomista del significado y prescindir de todo holismo. Sencillamente necesitaríamos disponer, junto a la gramática y al diccionario usual, de una especie de (monstruoso) diccionario de situaciones que determinara la aportación de cada expresión según el contexto. Una visión atomista del significado no precisa postular “entidades” que se correspondan con

estructuras gramaticales, cualquier postura que defienda la posibilidad de determinar *sistemáticamente* la aportación que una cierta estructura gramatical hace al significado de la oración en la que aparece, asume un cierto grado de atomismo.

3. Las paradojas semánticas

Pero el problema principal que Davidson debe afrontar para hacer un uso legítimo de la teoría tarskiana de la verdad, es aquel que originó la obra de Tarski: las paradojas semánticas. No basta con decir de ellas que se deben a un exceso de generosidad en el alcance de los cuantificadores. Como dice Tarski “el concepto de verdad (igual que otros conceptos semánticos) conduce inevitablemente a confusiones y contradicciones cuando se aplica al lenguaje coloquial en conjunción con las leyes normales de la lógica”⁸. La oración “esta oración es falsa” pertenece al lenguaje natural por mucho que uno alargue o acorte el rango de los cuantificadores. Una posible respuesta de Davidson sería afirmar que la incapacidad de construir un bicondicional (V) para la paradoja del mentiroso muestra que no es significativa. Pero esta solución es del todo ad hoc y no resulta convincente por las siguientes razones:

1) La oración del “veraz” (aquella que dice de sí misma que es verdadera)⁹ sí puede instanciar satisfactoriamente un bicondicional (V) y, sin embargo, es estructuralmente idéntica a la paradoja del mentiroso, ¿por qué una ha de tener significado y la otra no?.

2) No todas las paradojas dependen únicamente de la estructura de las oraciones en las que se instancian, existen paradojas que, a diferencia de la del mentiroso, dependen, además, del contexto. Cuando Epiménides el cretense dijo que “los cretenses mienten siempre” su preferencia podría haber sido falsa (nunca verdadera) si algún cretense hubiera dicho la verdad alguna vez, pero si todos, Epiménides incluido, hubieran mentido siempre, entonces el enunciado de Epiménides generaría una autodependencia enfermiza que desembocaría en paradojas¹⁰. ¿Carece por ello de significado la oración emitida por Epiménides? Obviamente no. ¿Carece de significado la *emisión*? Dado que Davidson se resiste prudentemente a hablar de enunciados como entidades, la pregunta sería ¿Hay oraciones que son significativas en ciertos contextos y no significativas

⁸ Tarski, 1956, pág. 267.

⁹ Para evitar problemas, podemos buscar formulaciones de la paradoja del mentiroso y de la oración del veraz que no tengan por sujeto un demostrativo, por ejemplo, (p)=“(p) es falsa” y (q)=“(q) es verdadera”.

¹⁰ Incluso oraciones como “Todo lo que Nixon dice sobre el Watergate es falso” dichas por Dean podrían ser paradójicas si entre las enunciaciones de Nixon sobre el Watergate encontráramos: “Todo lo que Dean dice sobre el Watergate es verdadero” (el ejemplo es de Kripke).

en otros? Yo creo que la respuesta es no, la enunciación tiene significado aunque sus condiciones de verdad no puedan satisfacerse. Nosotros no sabemos *a priori* si una enunciación es o no paradójica, éste es un hecho que, en casos como el de Epiménides, sólo constatamos después *empíricamente*, pero es preciso observar que nos resultaría del todo *imposible* establecer que una oración es paradójica si no fuera previamente *significativa*.

Descubrimos que la oración de Epiménides es paradójica cuando intentamos *satisfacer* los requisitos de su verdad y vemos que, dadas las circunstancias, es imposible. Lo que importa aquí es que disponemos de una serie de *prácticas regladas* determinadas en parte (pero sólo en parte) por cuestiones gramaticales, pragmáticas, etc. que traemos a colación al intentar evaluar la verdad de una oración. A veces estas reglas no son suficientes para evaluar veritativamente una oración, otras veces las reglas no son consistentes y, en muchas ocasiones, pertenecen incluso a sistemas distintos que circunstancialmente se solapan, pero es precisamente su existencia la que nos autoriza a decir de una oración que es significativa. Esta observación se hace extensible, no sólo a las paradojas empíricas o contextuales sino también a las que hemos llamado estructurales, de hecho, autores como Burge, Barwise y Etchemendy, Gaifman, o K. Simmons han ofrecido teorías en las que es posible evaluar semánticamente las paradojas atendiendo a criterios pragmáticos y contextuales.

3) Como Davidson señala, una de las razones por las que el concepto de verdad no es irrelevante es porque nos permite cuantificar sobre oraciones verdaderas, así pues, es de extrema importancia que la teoría que se nos propone sepa dar cuenta del significado de oraciones con cuantificación universal y existencial. A tal efecto, la convención (V) no basta porque entonces algunos enunciados que hemos mencionado no serían significativos. Para evitar inconsistencias necesitamos algo más que la convención (V), necesitamos la teoría tarskiana de la verdad al completo pero esto supondría una reforma del lenguaje natural intolerable. Para comenzar, necesitaríamos una jerarquía de lenguajes de capacidad expresiva creciente y esto no es posible cuando estamos intentando dar el significado del lenguaje natural en su conjunto a través del propio lenguaje natural, cuando metalenguaje y lenguaje objeto son una y la misma cosa.

Para finalizar y, a modo de conclusión, diré tan sólo que la reducción de la noción de significado a la de verdad en términos tarskianos tropieza con dificultades a mi juicio insalvables por todo lo expuesto hasta ahora. Por otra parte, el marco holista en el que seguramente debemos situar el significado quizá exceda el ámbito de las reglas puramente lingüísticas y dé cabida a

muchos otros sistemas de reglas que interfieren circunstancialmente en la determinación del significado. En uno de sus artículo “El pensamiento” Frege sugería que una definición general de la noción de verdad incurriría en una petición de principio porque las oraciones utilizadas en tal definición deberían ser previamente verdaderas. Quizá este tipo de paradojas ponga de relieve la dependencia mutua que se da entre significado y verdad, el vínculo es tan estrecho que, tal vez, todo intento de reducir completamente una noción a la otra resulte al fin y al cabo estéril.

BIBLIOGRAFÍA

Davidson, Donald (1984) *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford. (La traducción de las citas es mía).

Tarski, Alfred (1956) “The Concept of Truth in Formalized Languages”, en *Logic, Semantics and Metamathematics*. Clarendon Press, Oxford (La traducción de las citas es mía).

Jordi Valor Abad

Universitat de València

e-mail: Jordi.Valor@uv.es

SOBRE LA COMPRENSIÓN DE LOS NOMBRES PROPIOS

IGNACIO VICARIO ARJONA
Universidad de Barcelona

ABSTRACT: In the paper I will be concerned with the question of understanding proper names. I argue that the belief that forms a hearer who accepts as true a sentence that includes a name has to meet more requirements than just to get the correct object, insofar as the content of a speaker's belief that *a* is *F* may differ from the content of her belief that *b* is *F*, even if *a* is *b*. But I do not think that that commits us to include something like a sense in the meaning of names.

KEYWORDS: Proper name, understanding, reference, sense.

Cuando nos preguntamos por aquello que un hablante debe captar a fin de que pueda decirse que ha comprendido el uso de un nombre propio es de todo punto natural que lo primero que se nos venga a las mentes es que ese hablante debe llegar hasta lo que sea que designe el nombre, es decir, hasta el portador del nombre, el referente.

Esta posición es considerada a menudo insuficiente, aun cuando se acepta que captar el referente es un paso obligado en la comprensión de los nombres. En lo que sigue se mostrará que aunque es preciso calificar esa afirmación inicial, tal calificación no ha de llevarnos mucho más lejos.

Por lo común entender un nombre propio supone cierto conocimiento previo del objeto en cuestión por parte del oyente. Algunas situaciones son excepciones nada infrecuentes a esta normalidad. No parece que seamos reacios a aceptar que un oyente entiende lo que se le dice en ocasiones en las que no cabe suponerle conocimiento alguno acerca del referente. Por ejemplo, cuando damos un nombre para hacer una reserva en un restaurante, en un hotel, o para pedir hora al médico o en la peluquería. En estos casos parece suficiente que el oyente comprenda que se trata de alguna persona, quizá la propia que hace la petición de reserva, que tiene tal nombre. Estos usos de nombres se sostienen meramente en el conocimiento lingüístico que los hablantes poseen, en un sentido implícito, de algunos aspectos de su uso, relativos a la sintaxis y a su función semántica. Sin embargo, en la situación típica en que los nombres

proprios son utilizados, aquellas en que se sostiene su peculiaridad expresiva, se persigue un fin más preciso, como es el de llamar la atención del interlocutor acerca de cierto sujeto en concreto. A ese fin, se cuenta con que el nombre no sea nuevo para el oyente, y que éste conozca algún tipo de información acerca del referente. Tal cosa sucede cuando alguien nos dice de pronto “Siempre he admirado profundamente a Velázquez”; el hablante está suponiendo que su interlocutor posee algún tipo de noción acerca del portador de dicho nombre. Este fenómeno no se reduce a situaciones en las que intervienen nombres de personas ilustres o famosas, sino que el mecanismo se halla presente también en el caso en que unos padres hablando entre sí hacen referencia a su hija Elvira mentándola por medio de su nombre. Puestos a dictaminar si ha habido o no comprensión, en ninguna de estas situaciones, concebidas de modo apropiado, nos daríamos por satisfechos si el oyente se limitara a colegir que su interlocutor está haciendo referencia a alguna persona llamada “Elvira” o “Velázquez”, sin apreciar en ese mismo momento que esa persona es, en un caso, su propia hija o, en el otro, el genial pintor. Ciertamente no es muy urbano o desde luego muy conveniente, si uno espera hacerse entender, dejar caer en la conversación nombres cuando no se tiene la confianza de que nuestro interlocutor está en disposición de identificarlos por medio de algo más que ser, sencillamente, la persona o el objeto de que le están hablando a uno en ese momento. (Lo normal es que la utilización de un nombre que se supone nuevo para la audiencia vaya acompañada de una pizca de información que se pueda asociar al nombre.)

Si entender un nombre es dar con el referente, cualquiera sea el medio por el que el oyente llegue a concebir que el nombre está por el referente será suficiente para entender. Sin embargo, algunas situaciones parecen desmentir esta flexibilidad. Fácilmente se reconocerá que hay cosas que son nombradas por más de un nombre, es decir, que hay nombres que son distintos a pesar de que refieren a la misma cosa. Los pares de nombres “Héspero” y “Fósforo”, o “Tulio” y “Cicerón” son algunos de los más conocidos por los filósofos. Es claro que tanto si se dice “Tulio denunció a Catilina” como si se dice “Cicerón denunció a Catilina” se está asertando un mismo hecho del mundo. Sin embargo, no es tan claro que se requiera por parte del hablante exactamente lo mismo al entender una oración que al entender la otra. Alguien puede entender una, sin entender la otra; y, lo que es más, alguien puede entender ambas sin haber de reconocer por ello mismo que Tulio es Cicerón.

Esta circunstancia nos obliga, pues, a reconsiderar la posición inicial: la de que entender un nombre consistía en dar con su referente. Ahora bien, en principio no parece que sea preciso retocarla. Las diferencias de comprensión se deben a dos factores distintos y fácilmente explicables. El primero es que, evidentemente, estamos ante expresiones que difieren fonológica y ortográficamente. El reconocimiento de una no es el reconocimiento de la otra.

El segundo es que el conocimiento de objetos es esencialmente parcial. Los objetos presentan facetas diversas, lo que propicia que sean conocidos de varios modos hasta el punto de que una misma persona no se aperciba de que lo que hasta el momento ha venido tomando por dos objetos distintos son en realidad el mismo. Así pues, basta para explicar las diferencias de comprensión de un hablante respecto de dos nombres correferentes señalar que tales diferencias son el resultado de que dicho hablante ha aprendido a asociar con una de las expresiones un objeto al que reconoce bajo determinadas facetas y a asociar con la otra expresión el mismo objeto bajo otras facetas distintas, que él no sabe que pertenecen al mismo objeto.

El hecho de que el hablante asocie facetas diversas del objeto con cada uno de los nombres puede llevar a pensar que tales facetas son parte integrante del significado de los nombres; y que, por ejemplo, el nombre “Fósforo” referiría al planeta Venus cuando aparece por levante al alba, en cuyo caso erraría el hablante que aplicara el nombre cuando vislumbra a Venus al caer de una tarde. Nada impide que una semántica semejante sea el caso. Es posible que nuestras lenguas contengan algunos nombres para los que sea norma el ser aplicados en función de ciertos datos relativos al referente. Es probable que los nombres “Héspero” y “Fósforo” sean dos de tales nombres. De esta suerte de nombre puede decirse que su significado incluye además de su referente una serie de rasgos, aspectos o propiedades del referente que un hablante debe conocer a fin de ser competente en el uso del nombre. Sin embargo, no parece que los más de los nombres que a diario utilizamos funcionen de modo semejante. Aun así, en lo que sigue examinaremos alguna de las razones que se formulan a favor de que los nombres incluyan tales perfiles descriptivos.

Que un hablante que no sabe que Tulio es Cicerón asocie con el nombre “Tulio” unos datos acerca del referente distintos de los que asocia con el nombre “Cicerón” no muestra en principio nada acerca del significado de una y otra expresión, ya que puede tratarse de meras diferencias psicológicas ajenas a la semántica. Basta con apreciar que la misma situación puede producirse cuando un único nombre está en juego, como en el ejemplo de Saul Kripke sobre Paderewski, en el que un individuo, movido por su sospecha de que los políticos no tienen dotes musicales, cree equivocadamente que el primer ministro polaco Paderewski es otro que el pianista Paderewski, cuando en realidad se trata de la misma persona¹. Sin embargo, como señala Richard Heck², la circunstancia de que el conocimiento de objetos presente facetas y el hecho de que una persona pueda creer de un objeto que tiene cierta propiedad y a la vez creer de ese mismo objeto que no la tiene, porque no se da cuenta de que es el mismo objeto, añade algunas dificultades a la explicación del

¹ Vid. “A Puzzle about Belief”.

² Vid. “The Sense of Communication”.

fenómeno de la comprensión. Supongamos que un hablante que no sabe que Tulio es Cicerón al oír “Tulio denunció a Catilina” entiende por lo que concierne al nombre lo que normalmente entiende cuando oye “Cicerón”. Es decir, un hablante que no sabe que Tulio es Cicerón oye “Tulio denunció a Catilina” y entiende que Cicerón denunció a Catilina. No hay duda de que esa persona a pesar de su confusión ha dado con el referente correcto, pero no es menos cierto que ello ha sido por casualidad. Parece justo decir que ese hablante no ha entendido lo que se le ha dicho, pues ha confundido un nombre con otro, aunque curiosamente el hecho del mundo que se representa sea el mismo que el que ha sido aseverado. Las dudas que pueden surgir acerca de si el hablante en efecto ha dejado de entender lo que se le decía habrían de despejarse tan pronto se apreciara que difícilmente podría decirse de ese hablante que está justificado en la creencia que ha formado en respuesta a la aseveración; sin embargo, es de esperar que una de las metas de la comunicación sea la transmisión de conocimiento o cuando menos de creencia justificada, y no simplemente la transmisión de creencia verdadera.

Pero en qué consiste el fallo que se ha producido en la comprensión. Sin duda no puede ser achacado a la equivocación del referente, pues tanto “Tulio” como “Cicerón” tienen el mismo referente. Parece pues que lo que se ha producido es un error en el reconocimiento de la expresión. Pero no sólo ha ocurrido eso, y de hecho no conviene fijarse exclusivamente en las diferencias de los nombres en tanto que expresiones (es decir, en sus propiedades fonológicas u ortográficas). Es frecuente encontrar nombres distintos que, no obstante, comparten expresión o palabra. Por ejemplo, en el sitio de Troya hubo dos guerreros de nombre “Ayax”; y en la Gran Bretaña hay un Blair Primer Ministro y hubo un Blair escritor. La homonimia es un hecho corriente entre los nombres propios. Lo habitual, no obstante, es que esos nombres tengan referentes distintos. Sin embargo, puede imaginarse una situación en la que nombres distintos compartan expresión y referente. Siguiendo un ejemplo de Richard Heck³, tomemos al escritor Eric Blair, más conocido como George Orwell, y supongamos que es ingresado amnésico en un hospital. Incapaz el paciente de dar un nombre, los médicos del hospital deciden llamarlo “Orwell”, tal vez por admiración hacia el escritor, pero, eso sí, sin que se hayan dado cuenta de que su paciente es en realidad el escritor. A partir de la situación dibujada puede reproducirse lo que antes acontecía respecto de los nombres “Tulio” y “Cicerón”. Uno de los médicos puede oír “Orwell luchó en la Guerra Civil Española” a alguien que hace referencia al escritor y entender que se hace referencia a su paciente amnésico. Aunque da la casualidad de que el paciente es el escritor, parece que, similarmente a como ocurría antes, el médico no ha comprendido lo que se le decía, puesto que ha confundido un nombre con otro.

³ Ibid.

La única diferencia entre el caso de “Tulio”/“Cicerón” y el de “Orwell” es que en éste los nombres homónimos comparten expresión y en el otro no. Lo cual pone de manifiesto que el problema de explicar el fallo de comprensión que tiene lugar en los ejemplos excede el hecho de si los nombres homónimos que intervienen difieren en sus propiedades fonológicas u ortográficas.

Pero, ¿qué más hace falta para explicar ese fallo de comprensión? Lo que se nos pide es que se dé un criterio para distinguir esos nombres por algo más que por su referente y por sus propiedades en tanto que expresión. Si nos fijamos en la forma en que la situación ha sido presentada advertiremos que ciertas propiedades, ciertas facetas del referente parecen tener alguna intervención en lo sucedido. Apréciase que al exponer el caso de la aseveración que confunde al médico nos veíamos obligados a hacer mención de la faceta de Orwell como escritor y de su faceta como paciente amnésico. Precisamente en esos términos parece poder explicarse la confusión en que cae el médico: lo que sucede es que el médico cree que la aseveración hace referencia al individuo Orwell en tanto que paciente en lugar de hacer referencia al individuo Orwell en tanto que escritor; es decir, ha pensado en el referente del nombre proferido bajo las propiedades que atribuye al referente del otro nombre, el nombre hospitalario. Pero ¿por qué explica esto que el médico no haya entendido la aseveración?, ¿por qué que el médico piense en Orwell de un modo u otro explica que haya confundido los nombres? La respuesta que suele ser ofrecida por aquellos que, como Heck, resaltan la influencia en la semántica de la peculiaridad del pensamiento de involucrar facetas o modos de presentación de objetos es afirmar que los nombres tienen convencionalmente asociados conjuntos de propiedades del referente. Aplicado a nuestro caso, lo que afirman es que el seudónimo literario “Orwell” y el apodo hospitalario “Orwell” difieren, a pesar de coincidir en expresión y referencia, en que tienen algo así como sentidos distintos (por recurrir a un término muy conocido). Según esto, para entender un nombre un hablante habría de dar con el referente a través de alguna de las facetas contenidas en el conjunto de propiedades asociado con el nombre, así, por ejemplo, para entender el seudónimo “Orwell” habría, tal vez, de pensar en Orwell como autor de ciertos libros, y para entender el apodo hospitalario quizá habría de pensar en Orwell como el paciente amnésico de la habitación 627. Pero esta explicación da un paso demasiado grande que no parece venir exigido por el problema bajo consideración. No es suficiente para incluir en el significado tales conjuntos de propiedades llamar la atención sobre la circunstancia de que en la comprensión de uno de los nombres ese hablante se representa al referente bajo un modo de presentación distinto de aquel bajo el que se representa al referente cuando entiende el otro nombre. Por lo demás dicha propuesta deja muchos detalles sueltos que son de difícil explicación. No se dice, por ejemplo, cómo se constituye ese conjunto, qué determina las propiedades que lo integran, ni por

qué la consideración de una de ellas le permite a un hablante entender. Pero no abordaremos aquí la crítica a una concepción tal. En su lugar indicaremos brevemente otra línea de explicación, que acude a una noción conocida, como es la de una cadena de comunicación, y satisface la intuición de que el modo de pensar en el referente es pertinente en la comprensión, sin haber de admitir sentidos.

La pregunta a la que había que dar respuesta es: ¿qué diferencia un nombre de otro? Si hemos estado dando por sentado que el seudónimo "Orwell" y el apodo "Orwell" son nombres distintos es porque tienen orígenes bien diferenciados. Son el producto de dos convenciones separadas. Una, la que seguimos nosotros, es bastante más antigua que la otra y fue iniciada probablemente por el mismo Eric Blair. La otra se remonta al paso de éste por el hospital, y fue el producto del acuerdo de unos médicos. Las historias causales de una y otra son claramente distintas. Los nombres están sustentados en distintas cadenas de comunicación. Sobre estas cadenas diremos algo más adelante. De momento tratemos de centrarnos en la cuestión de la relación entre esta forma de individualizar nombres y el fenómeno de la comprensión. Pero antes que nada insistamos por un momento en la relación entre las personas y la información que almacenan acerca de otras personas. Dijimos que, en la situación normal en la que el nombre no es nuevo para un hablante, éste al comprender el uso de un nombre lo que hace es identificar el objeto que es el referente, en el sentido de que pone en relación el nombre que oye con la información que el hablante guarda de ese objeto. Algo parecido pasa también sin que intervenga el lenguaje. Cuando una persona identifica a otra con la que se cruza por la calle lo que sucede es que pone en relación la información perceptual que le llega en ese momento con la información que guarda de otros encuentros con esa persona. Para facilitar la exposición supongamos que la información que concierne a cada persona que conocemos está clasificada en carpetas. En cada una de ellas almacenamos los datos que percibimos o que nos cuentan del individuo al que está dedicada. Naturalmente, entre la información que guardamos en la carpeta que corresponde a una persona está su nombre, si es que lo sabemos. Así, cuando oímos un nombre que creemos identificar lo ponemos en relación con la carpeta que corresponde al que creemos es su portador.

Pero ¿qué determina que el nombre que aparece en una aseveración sea comprendido o no? Lo que determina la comprensión del nombre es la relación que la carpeta mantiene con ese nombre. Ahora bien, esa relación no consiste en que alguno de los datos de la carpeta coincida con los datos que supuestamente corresponderían al sentido del nombre. En realidad la relación en cuestión es de tipo causal. Lo que importa para que haya comprensión es el vínculo entre la carpeta y la cadena de comunicación, en la que se sostiene el

uso del nombre, que se establece cuando ejemplares anteriores del nombre proferido han sido identificados en relación con esa misma carpeta.

Cuando en la situación mencionada al principio uno de los padres que están charlando entre sí oye el nombre “Elvira”, el acto de comprensión pasa porque identifique el referente en el sentido de que relacione el nombre proferido con la carpeta donde guarda la información acerca de su hija, y no se limite meramente a entender que su cónyuge le está hablando de una persona llamada “Elvira”. Ahora bien, la cuestión fundamental no es que la carpeta guarde información relativa a su hija, pues hemos visto que es posible tener por confusión dos carpetas que correspondan a una misma persona; sino que lo que importa para la comprensión del nombre es que parte de esa información se haya obtenido a partir de usos anteriores de ese mismo nombre; dicho de otro modo, lo relevante es que típicamente usos anteriores del mismo nombre hayan ido a parar a esa carpeta. A quien cree que Tulio es otro que Cicerón le sucede que tiene carpetas distintas (sin relacionar) para la información que atribuye a Tulio y para la información que cree corresponde a Cicerón. Y a cada una de estas carpetas van a parar respectivamente los usos del nombre “Tulio” y los usos del nombre “Cicerón”. Si ese hablante en una ocasión asocia una preferencia del nombre “Tulio” con la carpeta con la que ha asociado hasta el momento los usos del nombre “Cicerón” entonces no comprende, a pesar de que esa carpeta guarde información del mismo objeto que el referente del nombre. Lo que ha sucedido es que ha dado con el referente de otro nombre, que por casualidad es el mismo que el del nombre proferido. De modo similar ocurre con los médicos del hospital respecto de Orwell: tienen dos carpetas distintas para la misma persona. Una de ellas se ha formado en parte con información proveniente de usos del seudónimo “Orwell” y la otra se ha formado en parte con información proveniente de usos del apodo hospitalario “Orwell”. Cuando uno de esos médicos asocia un uso del seudónimo literario con la carpeta que ha formado en relación con los usos del apodo hospitalario entonces no ha comprendido, a pesar de que, nuevamente, ambas carpetas corresponden al mismo individuo y a pesar de que en ambas consta como nombre “Orwell”.

Es cierto que si puede producirse un error de identificación de un nombre también pueden producirse muchos. Esas carpetas pueden contener datos obtenidos en relación con otras cadenas de comunicación, producto de algunas confusiones como las que estamos describiendo. Naturalmente, si la confusión se generaliza y a una misma carpeta, en la que consta un solo nombre para su portador, van a parar las preferencias de dos nombres distintos, tal vez homónimos, entonces no hay manera de determinar si la carpeta se corresponde con un nombre o con el otro, y lo que debe decirse es que ese hablante no comprende ninguno. Es una difícil cuestión establecer la proporción de errores que una carpeta puede soportar al ser asociado con un

nombre, pero no es una dificultad propia del lenguaje, sino que tiene que ver mayormente con el pensamiento y el problema de determinar su contenido. Puede suceder, también, sin que ello cree ningún tipo de problema, que una carpeta contenga más de un nombre, como ocurrirá con todos aquellos hablantes que saben que Tulio es Cicerón o que Orwell es Blair.

A fin de calibrar adecuadamente las peculiaridades del presente análisis conviene que nos fijémonos en lo siguiente. Dependiendo de la información que un hablante guarde en una carpeta variará su modo de concebir al individuo al que corresponde la carpeta, es decir, variará el modo en que se le presenta ese individuo en el pensamiento. Pero la manera en que se ha explicado el hecho de que haya o no comprensión no depende de cuál sean los datos contenidos en esas carpetas. No importa qué datos acerca del ilustre romano guarda en cada carpeta un hablante que desconoce que Tulio es Cicerón. De hecho, puede ocurrir que este hablante guarde en la carpeta para Tulio algunos de los datos que otro hablante semejante guarda en la carpeta para Cicerón. Es más, puede suceder que todo lo que uno atribuye a Cicerón el otro se lo atribuya a Tulio y viceversa. Todo menos el nombre, claro, pero incluso respecto a eso puede no haber diferencias como se muestra en el caso de los dos nombres "Orwell". Lo cual, sin duda, pone en aprietos a la estrategia que dota a los nombres de perfiles descriptivos.

El esbozo de análisis que se ha expuesto se basa en dos elementos principales: por un lado, el hecho de que ejemplares diversos de cierta expresión sean ejemplares de un mismo nombre si pertenecen a una misma cadena de comunicación; y por el otro, el hecho de que una carpeta haya sido constituida en parte con información obtenida en relación con una misma cadena de comunicación. Sobre el primero de ellos apenas se ha dicho nada y bueno será que acabemos defendiéndolo ante alguna de las críticas que pueden hacersele.

La idea de una cadena de comunicación persigue básicamente explicar cómo se preserva la referencia entre los usos últimos de un nombre y los primeros; es decir, que la cadena lo que logra es establecer una unidad, una trabazón entre los usos del nombre que van efectuando los hablantes. En *Naming and Necessity*, Saul Kripke dispuso que una vez entraba en circulación el nombre, tras ser fijada su referencia en el acto de nominación o "bautizo", se preservaba el referente de los usos posteriores del nombre mediante una cadena de dependencias entre hablantes, cuya eficacia requería únicamente que el nuevo hablante tuviera la intención de seguir el uso establecido y que él no haría sino heredar.

La simplicidad de esta idea fue rápidamente puesta en entredicho por la posibilidad, señalada por Gareth Evans⁴, de cadenas de comunicación que se

⁴ Vid. "The causal Theory of Names".

desvían y pasan a referir a otro objeto sin que se viole el requisito de Kripke, es decir, sin que nadie en la cadena haya albergado la intención de cambiar la orientación del uso establecido. Tal cosa aconteció en realidad a propósito del nombre “Madagascar”. Este nombre refería, al parecer, a una parte del continente africano cuando Marco Polo lo tomó de marinos árabes o malayos y, por confusión, lo difundió en Europa como refiriendo a la gran isla africana que hoy designa. El propio Evans, sin embargo, explicaría posteriormente⁵ cómo remendar la noción de una cadena de comunicación. Las cadenas habrían de entenderse como prácticas que se van haciendo: en ellas la referencia no se fijaría de una vez en un único acto, el “bautismo”, sino a través de toda una serie de situaciones en las que el objeto sería identificado por otro medio que como el referente del nombre y en las que se diría su nombre, como ocurre, por ejemplo, cuando se introduce a un hablante en el uso de un nombre diciendo, a la vez que se señala al referente, “Ésta es Julia”.

Llegados a este punto, el teórico que defiende incluir sentidos en los nombres puede replicar que el hecho de que una cadena de comunicación gire en torno a un objeto u otro depende de esas identificaciones del objeto que llevan a cabo los hablantes y que sirven para fijar la referencia; identificaciones que se apoyan en último término en las propiedades con la que los hablantes distinguen ese objeto del resto de objetos. El caso de Orwell ilustra especialmente la dificultad planteada. La existencia de dos cadenas de comunicación que se sirven de la misma palabra y que además giran en derredor de un mismo objeto sólo puede producirse si esas cadenas están nítidamente separadas. Es decir, depende de la coherencia con la que los médicos llevan a cabo las identificaciones del paciente en tanto que referente de una de las dos prácticas. Si las dos cadenas de comunicación subsisten sin mezcla, salvo por alguna confusión ocasional, es por la homogeneidad de los criterios con los que esos médicos distinguen entre una faceta y otra de Orwell.

El razonamiento es atractivo, y recupera toda la fuerza sugeridora que está detrás de aquellos ejemplos en que un objeto presenta para los hablantes dos facetas claramente separadas, como sucede con Héspero y Fósforo, o con Superman y Clark Kent. Pero tiene razón en un sólo punto: en algunos casos, como es el de Orwell, la existencia de esas facetas separables explica la circunstancia de que puedan preservarse durante algún tiempo sin formar un batiburrillo incoherente dos cadenas de comunicación erigidas sobre una misma expresión y sobre un mismo objeto. Ahora bien, el razonamiento no logra su objetivo por cuanto ignora qué es lo verdaderamente esencial de esos actos identificadores en los que se fija la referencia y sobre los que se sustenta la cadena de comunicación. Y es que la existencia de una cadena de comunicación depende únicamente de que un objeto en concreto sea identificado. Esto sólo

⁵ Vid. *The Varieties of Reference*.

exige que haya algún modo de llevar a cabo esas identificaciones, pero no requiere que deba haber alguno en particular. Antes bien, lo normal es que haya varios, incompletos, algunos bastante imprecisos, y que su número o validez se vean alterados durante la vida del nombre. Así pues, aunque puede suceder que algunos hablantes compartan algún procedimiento para reconocer el objeto, no es algo que deba ocurrir en general. No forman parte de la cadena de comunicación, no son parte de lo que se preserva a través de la vida de un nombre, los criterios de identificación que utilizan efectivamente los hablantes en esos episodios que contribuyen a fijar la referencia del nombre. En realidad, cuáles sean en último término esos criterios seguidos en la identificación es algo que propiamente concierne únicamente a ese otro acto de referencia que tiene lugar en tales episodios, como cuando se dice “Ahí va Juan” o “Ésta es Julia”, y no al tipo de referencia que un hablante realiza mediante la proferencia de un nombre. Una vez que la práctica se ha cimentado a través de esos episodios identificadores en torno a un objeto un hablante puede seguir la práctica sin que ello suponga el conocimiento de los criterios con los que se estableció el uso del nombre, ya sea en el “bautizo” o en episodios posteriores. Fijémonos en que todos los hablantes actuales de una práctica antigua pueden continuarla, a pesar de que ignoren cómo se forjó, y a pesar de que la poca información, supuestamente de su portador, llegada hasta ellos puede resultar, a la postre, falsa. Si “Homero” refiere a un individuo en concreto no se debe a la circunstancia de que éste escribiera la *Iliada* y la *Odisea*, sino al hecho de que nuestro uso derive del uso de los que introdujeron el nombre para cierto sujeto en particular, aunque ésta nada tuviera que ver con esas obras. Es la existencia de un rastro lo que justifica el uso, más que las propiedades que se han ido adhiriendo por el camino. Y el rastro que configura la cadena de comunicación lleva hasta el objeto, descansa en éste, y no en los criterios por medio de los cuales fue seleccionado. Los nombres propios no están atados a conjunto de propiedades o procedimiento identificatorio ninguno. El modo en que somos introducidos en su uso no es más que un modo de introducción y no un elemento de la práctica misma. Hay una gran distancia, que no debe pasarse por alto, entre una práctica que consiste en nombrar la cosa que satisface cierto perfil y la práctica de nombrar cierta cosa, que nos es indicada gracias al hecho de que satisface cierto perfil, pero en la inteligencia de que es la cosa lo que cuenta, cualesquiera sean en adelante los procedimientos utilizados para establecer que se trata de la misma cosa.⁶

⁶ Las ideas aquí presentadas se han beneficiado de largas y fructíferas discusiones mantenidas con Manuel García-Carpintero. Asimismo, agradezco a José A. Díez, Joan Pagès y otros miembros del Grupo de Epistemología y Análisis sus comentarios a estadios previos de este trabajo.

BIBLIOGRAFÍA

- Evans, Gareth: "The Causal Theory of Names", *Aristotelian Society Supplementary Volume*, 47, 1973, pp. 187-208.
- Evans, Gareth: *The Varieties of Reference*, Nueva York, Oxford U. P., 1982.
- Heck, Richard: "The Sense of Communication", *Mind*, vol. 104, 1995, pp. 79-106.
- Kripke, Saul: "A Puzzle about Belief", en *Meaning and Use*, Margalit, A., ed., Dordrecht, Reidel, 1979, pp. 239-83.
- Kripke, Saul: *Naming and Necessity*, Oxford, Blackwell, 1980.

Ignacio Vicario Arjona
Universidad de Barcelona
e-mail: vicario@trivium.gh.ub.es





A RADICAL INTERPRETER'S PARADOX

ANTÓNIO ZILHÃO
Universidade de Lisboa

ABSTRACT: The main thesis characterizing the standpoint of interpretation is that empirical semantics has two possibility conditions - there has to be an essential link connecting speech and action and there must be a common behaviour of mankind. Contrary to other interpretationists, Davidson develops a concrete program of empirical semantics. He therefore needs to put forth a specification of the common behaviour of mankind. I think that the way he does it conflicts with the other interpretationist requirement that there must be an essential link connecting speech and action. Therefore, I contend that, if interpretationism is right, Davidson's program must fail.

KEYWORDS: Agency, bayesian decision theory, empirical semantics, interpretation, psychology of decision, rationality.

The so-called 'standpoint of interpretation' admits being characterized in terms of the thesis that empirical semantics has two possibility conditions - there has to be an essential link connecting speech and action and there must be a behavioural core common to interpreter and interpretee. As far as I know, Wittgenstein was the first philosopher to have explicitly formulated this standpoint. His statement: "The common behaviour of mankind is the system of reference by means of which we interpret an unknown language" (PI, §206) could easily be picked up as the slogan of interpretationism. Indeed, both Quine's program of radical translation and Davidson's program of radical interpretation are but developments of this basic idea of Wittgenstein's.

It is however one thing to state that the work of interpretation has to proceed having the common behaviour of mankind as its system of reference and another quite different to specify that behaviour in some theoretically illuminating way. That is, a way that does not get reduced either to a collection of examples of behaviours that trigger a feeling of empathy in us or to some paraphrase or other of the indexical characterization 'it is the behaviour of these creatures' together with some sort of ostension having the set of human

beings as its target.

The great *prima facie* theoretical advantage of Davidson's work over Wittgenstein's or Quine's resides precisely in the fact that, contrary to the latter's, it endeavours to offer such a specification. It rests on two moves. Taking seriously the old aristotelian definition of Man as the rational animal, on the one hand, and presenting an independent characterization of rational agency, on the other hand. Such a presentation permits of the subsequent identification of the common behaviour of mankind with the behaviour of rational agents. Therefore, what is taken to be the appropriate characterization of the latter is assumed to provide the system of reference by means of which completely unknown languages are to be interpreted.

Although I acknowledge the theoretical appeal of Davidson's strategy, I contend that his characterization of the common behaviour of mankind is not compatible with the other interpretationist requirement that there must be an essential link connecting speech and action. From such an incompatibility, it follows that, if the standpoint of interpretation is right, no effective program of empirical semantics can be undertaken within Davidson's theoretical framework. The arguments for this contention will be specified below.

Davidson's view of rational action is basically the one to be found in bayesian decision theory. It can be summarized thus: an action is rational if and only if it is such that it produces an outcome which is the best according to the agent's desires and the beliefs he entertains; an agent is rational if and only if his actions are, in general, rational. To be sure, such a definition does not yet yield a criterion to decide between competing courses of action which of them is the best; it is only a definition stating what it means to say of a given action that it is rational. In particular, it is a definition equating rationality with optimality. The criterion specifying how is one to assess which of a given set of possible actions is, under each circumstance, the best or one of the best is the so-called principle of 'maximization of expected utility'.

This principle can in turn be accounted for thus. Desires are meant to be pro-attitudes towards representations of outcomes of actions. The different outcomes taken to be possible under each particular set of circumstances are, in turn, supposed to be ranked in a scale determined by the assignment of real-numerical values to them, according to their respective degree of desirability or utility; this is the so-called 'utility function'. It is thus assumed that each agent has a well-defined utility function. Beliefs are meant to be truth-evaluable attitudes towards representations of particular states of the world assumed to be relevant to the determination of the outcome of the agent's action. The likelihood the agent ascribes to the obtaining of any of these particular states of the world varies along a probability scale. To act rationally in a particular situation is then to act in a way that admits being modelled as resulting from the following succession of steps. First, the subjective probability ascribed to

the obtaining of each relevant state of the world is multiplied by the subjective utility ascribed to the obtaining of each of the possible outcomes of the action. Secondly, the products so obtained are added; the sum thus obtained is then called the subjective expected utility of undertaking a given course of action. Thirdly, once steps one and two were iterated in connection with all possible courses of action, a decision to act in a certain way is reached by selecting that course of action the expected utility of which is maximal.

Now, human actions are part of our experience. They are events taking place in space and time before us. If a theory of rational action such as the one described above is to be taken to be a theory of human action, then its truth ought to be empirically verifiable, at least in principle. *Prima facie*, the decision-theoretical framework should admit being put to the trial of experience in the usual way. That is, the principle of maximization of expected utility should be taken as a law relating sets of upstream conditions with sets of downstream consequences. The theory would then be in agreement with reality if and only if upstream conditions and downstream consequences would relate to each other in the way prescribed by the lawlike principle. The verification of this requirement should then involve the following stages: first, the correct identification of the beliefs and desires, and their respective strengths, of the agent and of the action related to them in any particular occasion; second, the confirmation that agents *do* act in agreement with the principle of maximization of expected utility.

However, and as a number of authors have noted, one is faced with a major problem here. It is the following. The procedures set up within the theoretical framework of bayesian decision theory in order to allow us to arrive at a positive identification of the agent's beliefs and desires are procedures that make sense only if one assumes from the outset that agents act according to the principle of maximization of expected utility. The situation one is faced with is thus the following. In order to test the validity of a relational principle one needs, *prima facie*, to have an independent access to the *relata* such a principle relates. However, when one considers bayesian decision theory carefully, one realizes that, even assuming the content of human actions to be transparent, the assumption that the relational principle holds is a necessary condition to identify the upstream *relata*, namely, beliefs and desires. In other words, there seems to exist no access to a substantial part of the relevant empirical evidence outside of the framework of the theory itself. This means that no empirical check can be made before such a framework is itself subject to a closer examination.

What such a closer examination reveals is that bayesian decision theory is based on an analysis of what one might call 'the behaviour of a rational gambler'. That is, it is based on a couple of axioms that formalize the behaviour a gambler needs to have in a game of chance if he is to avoid loosing money

whatever happens. The thesis according to which this theory is empirically true boils down then to an analogical claim. Namely, the claim that human agents behave in their normal life in the same way as a rational gambler behaves in a game of chance. Now, does such a claim hold true of the facts? This is what needs to be checked.

Obviously, the fact that in a game of chance one needs to behave according to the axioms of the theory in order to avoid losing money whatever happens is by itself no proof that bayesian decision theory is empirically true of human behaviour in general. It may well be the case, for all we know, that most humans are not rational gamblers anyway and that therefore in a game of chance they would easily fall in situations in which they would lose to a professional gambler no matter what happens. Or it may well be the case that despite the fact that most humans behave as rational gamblers when placed in the context of a game of chance, they do not behave in real life as if it were a game of chance, i.e., the analogical claim may be false. Therefore, if the theory is to be empirically true of human actions in general then its axioms will have to hold in most everyday life situations. This is the crucial point. As a matter of fact, if the axioms hold true of the way everyday decision problems are solved by most people, then the validity of the principle of maximization of expected utility follows by purely mathematical means. That is, such a principle is a *theorem* of the theory and not one of its axioms.

Different axiom systems and slightly different versions for what are basically the same axioms have been put forth in the relevant literature. However, two of them are, in any of the available versions, crucial for the outlining of the structure of a rational gambler's behaviour. They are the following. First, the axiom I will call 'Axiom A'; it holds that a certain relation, namely, the relation 'it is at least as preferred as' (or, in another version, 'it is at most as preferred as') obtains between the coordinates of any pair of outcomes open to the agent's consideration and that such a relation is a weak linear ordering (or connected preorder), i.e., that it is a binary relation satisfying the following properties: transitivity, reflexivity and also connectedness. Second, the axiom I will call 'Axiom B'; it states that if an option A is at least as preferred as an option B then, if options C and D result from, respectively, options A and B by the introduction of the same change in their common outcomes, then option C is at least as preferred as option D.

Axiom A can be justified thus. It is this axiom that allows utilities to be put in a one-one correlation with real numbers, that is, that allows them to be measured in such a way that both their places in the utility ranking and their intrinsic differences of value are adequately expressed by the assignment of real-numerical representatives to them. Actually, transitivity is the essential property for this purpose. As a matter of fact, the admission that failures of

transitivity might exist would entail two consequences. First, the consequence that actual utility rankings are not representable by means of ordinal utility scales and that therefore the whole process of selecting that course of action the expected utility of which is highest becomes pointless. Second, the consequence that the holding of an intransitive gambling pattern is liable to make a gambler lose money for no gain whatsoever to a smart gambler, that is, the consequence that such a gambler is not rational in the decision-theoretical sense of this word.

Axiom B can be justified thus. It follows, together with other uncontroversial assumptions, from what is taken to be a very simple and intuitively appealing decision strategy, sometimes called the 'sure-thing principle'. Such a strategy is the following: when trying to figure out what is the act which is more advantageous to him, the agent should consider only those elements of a set of mutually exclusive and jointly exhaustive possible futures states of the world which have different outcomes (that is, states with the same outcomes should cancel out).

Both these axioms have already been put to test. The empirical adequacy of axiom A has been checked by experiments devised by Tversky. The empirical adequacy of axiom B has been checked and subject to extended discussions in connection with the so-called 'Allais' Problem'.

Tversky's experiments elaborate upon Condorcet's voting paradox. As is well known, Condorcet pointed out that social inconsistencies can arise out of consistent individual choices. In particular, he has shown that majority voting generates an intransitive pattern of preferences whenever a set of three individuals a, b and c consider a set of three options A, B and C in such a way that a ranks the options according to the scale ABC, b ranks them according to the scale BCA and c ranks them according to the scale CAB and the options are subject to a sequence of three alternative votings. Examples of this kind were later taken to provide a useful illustration of the unsolved difficulties associated with the idea of providing a social welfare function. The peculiarity of Tversky's own work consists in the fact that he has shown by means of a variety of psychological experiments that a similar pattern of intransitive preferences can arise out of the choices of a single agent who ranks his options on two dimensions in such a way that the agent's payoff matrix has the structure of a lexicographic semiorder. He has moreover shown that in a number of situations it is not uncommon for agents so to rank their options.

Allais' problem was initially introduced as a thought-experiment; however, it received later empirical confirmation. It consists in the successive consideration of two decision situations each involving two gambles. The result of the experiment is such that the set of two successive choices made by most agents violates Axiom B. These choices have the following character. In the first choice, agents typically choose a gamble in which they win a large

amount of money outright in place of a gamble in which the winning of that very same amount of money is uncertain and has to be weighed against a small probability of winning 5 times as much and an even smaller probability of winning nothing. In the second choice, agents are confronted with a set of two gambles in which the dominant outcome is winning nothing. In this case, they typically choose the gamble in which the highest prize is at stake, despite the fact that the probability of winning it is slightly inferior to the probability of winning the also large but yet 5 times smaller prize they chose in the first choice.

These and other similar experiments should have settled the question concerning the empirical validity of the theory. As a matter of fact, if the descriptive character of the theory hinges upon the descriptive character of its axioms, then, if such axioms are shown not to be empirically true of the behaviour of normal human agents in a non negligible number of decision situations, then it should follow that the theory as a whole cannot be ascribed the character of a descriptive theory of human choice behaviour. In particular, it should follow that its main theorem, namely, the theorem according to which agents maximize expected utility, cannot be ascribed the character of a lawlike principle relating the upstream conditions and the downstream consequences associated with the unfolding of human actions. And if this is so, it should also obviously follow that such a theorem cannot be simply assumed to hold in order to infer the content of an agent's beliefs and desires from such an assumption together with the knowledge of the content of some downstream effects and a bit of mathematical reasoning. Assuming that one has some independent way of identifying one's beliefs and desires, the principle codified in this theorem may or may not be a powerful normative standard for decision-making, prescribing the way one should go about making one's decisions in order to optimize one's gains; as an empirical proposition, however, it seems that it must be deemed to be false.

This conclusion is far from being universally accepted though. The verdict of empirical falsity has been evaded by quite a number of authors, a leading representative of which is precisely Davidson. The main argument underlying such a response is basically the following.

In order to test the above mentioned axioms, experiments had to be made. Now, what these experiments yielded were behavioural *data*, and these had to be interpreted. In particular, the actual choice behaviours of the agents had to be classified under particular characterizations. Otherwise, it would not have been possible to assert that, e.g., failures of transitivity had taken place. However, the above mentioned assumption according to which the content of our actions could be considered to be basically transparent is untrue. That is, the identification of the downstream relata depends as much as the identification of the upstream relata on the use of a theoretical framework

within which particular interpretations may be considered to make sense. And there is no compelling reason why the interpretational constraints that should be at work here should not be precisely those constraints which are set up by the very same axioms the experiments in question were trying to check. Under such circumstances, instead of providing evidence in favour of the inappropriateness of the tested axioms, the results of the experiments should be reinterpreted in such a way that the content of the actions displayed in them might come out agreeing with the axioms after all. And this is something that is not particularly difficult to do. Consider for instance the following reinterpretation proposals.

Take the case of Tversky's experiments first. Someone might challenge his conclusions by making the following claims. In order to uncover the alleged existence of intransitive sets of preferences the experimenter must confront different sets of choices of the same subject with each other. Now, by necessity, the subjects go through these choice processes in time. Therefore, each choice must be separated from any other choice by a certain amount of time. Under such circumstances, why should the experimenter not simply interpret the *data* as exhibiting the plain fact that under the sort of circumstances present in Tversky's experiments subjects tend to change their minds?

Take now the case of Allais' Problem. The correctness of the verdict that Axiom B has been violated in the experiment devised by Allais depends entirely upon an unformulated assumption. Namely, the assumption that the consequences under consideration in such an experiment are to be fully characterized by their monetary values. But why should they? Why not include non-monetary considerations, such as, e.g., considerations of risk-aversion, in the way the consequences should be perceived? The inclusion of considerations of such a sort may make it possible to reinterpret the *data* in such a way that Axiom B comes out of the experiments unscathed.

In sum, the idea is that the axioms of the theory should be taken to have the *status* of synthetic *a priori* truths about our behavioural patterns. Such a strategy can certainly not be ruled out from the set of *prima facie* admissible interpretive strategies. And given that there is no external court of appeal capable of judge once and for all the correctness or incorrectness of any particular interpretation in any particular situation, one must settle for fruitfulness in the long run. In the meantime, one should stick to the sort of reasoning by analogy first introduced by Ramsey, the outcome of which is a theory possessing two important advantages: it is conceptually simple and it lends itself to mathematical formalization.

I think that this line of reasoning runs the risk of leaving the doors wide open to the generation of all sorts of epicycles. As a matter of fact, I do think that, at least in the case of Tversky's experiments, the interpretation according

to which subjects developed intransitive sets of preferences is not only much more intuitively plausible and charitable than the rival strategy of ascribing a highly flickering character to quite a number of human subjects but it also makes more cognitive sense. However, I am not going to pursue this line of argument here, as I have already done it elsewhere. What I want to stress here is the following aspect. I think that the line of escape that appeals to the synthetic *a priori*, implausible as it may be, is nonetheless a possible line of escape for decision-theorists such as Ramsey, Savage or Jeffrey. I also think that such a line of escape is not available to the theorist who made an explicit appeal to it, namely, Davidson. Below I will explain why.

Both Ramsey and Savage used gambles as the means to find out what agents believed and desired and how seriously did they do it. It was a consequence of their proposal that the utility scales and the probability distributions ascribed to agents might be strongly at variance with the content of the sentences of the very same agents concerning the objects of their desires and the contents of their beliefs. Indeed, the whole idea of the enterprise was precisely to create the theoretical framework within which a scientific study of human actions would be made possible. As in other fields of study, it was only to be expected that the result of the theoretic activity would frequently be in disagreement with people's impressionistic and intuitive assessments of the phenomena under consideration. Obviously, it is in natural discourse that such an impressionistic and intuitive assessment of psychological phenomena gets expressed. Therefore, no wide agreement was to be expected between either first-person or third-person natural discourse descriptions of intentional behaviour and the reports of the outcome of theoretically conducted behaviour experiments. As Ramsey put it, his 'artificial system of psychology' was to be compared with Newtonian mechanics and not with what one might term 'Folk-Physics'. Thus, Ramsey's identifications of the belief-desire structures of agents were meant to be theoretical identifications of the causes of actions, of the same sort of, e.g., the identification of water with H₂O, and not conceptual analyses of the way agents usually express their views about the unfolding of actions. Natural discourse played a role in the application of this 'artificial system of psychology' only insofar as subjects had to understand the gambles they were invited to consider.

The objects to which Ramsey and Savage ascribed degrees of belief were possible states of the world and the objects to which they ascribed degrees of desire were consequences of acts. The connection between them was established through the above-mentioned gambles. Jeffrey, however, replaced propositions for both possible states of the world and consequences of acts as unified objects of belief and desire. He understood propositions as abstract objects referred to by sentences but in no way exchangeable with them. And, in particular, he considered that the sentences proffered by

subjects concerning the objects of their own beliefs and desires were only in an evidential relation rather than in a relation of reference with the propositions that indeed constituted the objects of the subjects' beliefs and desires. He clearly endorsed the thesis that one may be in doubt or in error about what one's beliefs and desires are and that the touchstone is actual choice. In particular, he stated that agents may be unwilling to assent to the very same sentences that are used by the theorist to correctly express the proposition that is the content of their belief or of their desire. Thus, he also thought that the outcomes of the scientific study of action might be at variance with what natural discourse has to say on such a topic. The criteria for belief and desire are, according to him, behavioural, and speech is but a particular sort of behaviour among others.

All this is in perfect harmony with the idea according to which the axioms of bayesian decision theory are synthetic a priori truths about the structure of human intentional behaviour. One may suspect that there is no more truth in such a contention than there is truth in the contention that the axioms of euclidian geometry are synthetic a priori truths regulating our geometrical knowledge but that is all. Davidson's Unified Theory of Interpretation is, however, a different matter.

This theory contains a slightly modified version of Jeffrey's version of bayesian decision theory, namely, a version in which Davidson substitutes uninterpreted sentences for propositions as the objects of preferences. Such a substitution is needed, because Davidson uses decision theory not as the framework of a scientific theory of action but rather as the scaffold by means of the support of which the work of radical interpretation is supposed to succeed. Now, we have seen above that, according to what seems to be our intuitive way of classifying it, the behaviour of the members of our own community is frequently at variance with the expectations decision-theory brings about. It takes an extra interpretive effort, supported by a lot of theoretical argument carried out *in* the languages we already speak and understand, to accomodate that behaviour, under those circumstances, within the framework laid down by the theory. It may well be that the outcome of such work of accomodation is the right scientific or para-scientific account of such behaviours. But if this is true, then what normal people tend to say about such behaviours, even assuming their sincerity, is more often than not wrong. Therefore, the first of the above mentioned possibility conditions for empirical semantics, namely, the requirement that there be an essential link connecting speaking and acting, is not satisfied. If this condition is not satisfied then, if we assume the thesis of interpretationism to be correct, no empirical semantics is possible.

Let me elaborate a bit more upon this. Imagine that some choice behaviours are being performed by members of foreign communities the

language of which we completely ignore. If we are going to use the axioms of decision-theory as the scaffold in terms of the support of which we are going to interpret not only their actions but also what they say about them, then, if we assume, with Davidson, that the conceptual system underlying the semantics of their unknown natural language is going to be similar to ours, then their common way of describing their common behaviour is going to be as wrong as ours; but if this is so, then the common behaviour of mankind, as formally characterized in the decision-theoretical apparatus, cannot be the system of reference by means of which their unknown natural language is interpretable. Note that the only way of avoiding this conclusion is to postulate that the speakers of this unknown foreign language are inborn scientific psychologists. And remember that, assuming that some version or other of Ramsey's 'artificial system of Psychology' is true, we *know* that we are no such thing. This seems therefore a preposterous postulate to adhere to.

The introduction of Davidson's famous postulate according to which most of our beliefs are true can actually be accommodated within this state of affairs. To see this, we need only remember that, within Ramsey's 'artificial system of psychology', beliefs are defined as the (partial) causes of action. No assumption of transparency is made. Therefore, an agent's beliefs, *qua* partial causes of action, may well be true, and yet, still assuming sincerity, the sentences in terms of which he, or somebody else, formulates them may well be false, at least in the case of those belief-sentences which are not immediately prompted as a response to perceptive input. That is, it may well be the case that the agent (or his untrained observer) does not know or is not aware of what beliefs he in fact has. Again, that something like this should be expected to be the case was precisely the scientific motivation underlying Ramsey's development of his 'artificial system of Psychology'. Introducing such a postulate does not therefore necessarily commits one to the interpretationist idea that there is an essential link connecting speech and action.

Therefore, a supporter of interpretationism will have to criticize Davidson's approach in one of the two following ways. Either he challenges the cogency of Davidson's independent characterization of rational agency, or he challenges the view according to which Man is the rational animal. Given the vagueness of the concept of rationality itself, the choice between these two criticisms gets eventually reduced to a mere choice of words.

As a matter of fact, if he enjoys the use of the term 'is rational' in association with human matters then he should want to stick to the old aristotelian definition of Man. Under those circumstances, the interpretationist should criticize the allegedly independent definition of rational agency Davidson puts forth as wrong. But if he does not have any strong feelings associated with the use of the predicate 'is rational' to characterize the ways of

Man, the interpretationist could still easily accept Davidson's definition of rational agency as correct and at the same time deny the rationality of Man. Under such circumstances, he would have to refuse the thesis that a definition of rational agency ought to be used as the system of reference by means of which any unknown language is to be interpreted and he would have to look for a less grand and more parochial way of characterizing the ways of Man. Either way, what is relevant is that Davidson's characterization of rational agency turns out not to be an appropriate way of modelling the manner in which human agency tends to be characterized in our natural languages. Thus, grounding the work of interpretation upon such a characterization is most certainly going to lead the interpreter into the selection, as interpretive, of non-interpretive T-theories for the natural languages of foreign communities.

Of course, the option of accepting the decision-theoretical definition of the rational agent as basically true of us and of considering that scientific psychology should be built upon it is always open. But such an option comes with a price. Namely, the price of severing the supposedly essential link connecting our common ways of speaking and our common ways of acting. Severing that link has two clear consequences. The first is that ordinary ways of talking will have to be dismissed as irrelevant for the understanding of our common ways of acting; the second is that the knowledge of our common ways of acting will have to be dismissed as irrelevant for the understanding of our common ways of talking. Or, at least, they will have to be so dismissed until somebody comes along with a theory explaining in detail the convoluted ways in which our untrue ordinary way of talking about our actions is related to the true explanatory theory of action and vice-versa. Such a theory does not seem to be in the offing, however.

In closing, I would like to say that one cannot have it both ways: to claim, in the hope of procuring a foundation for empirical semantics, that the essential connection between speaking and acting should be preserved, on the one hand, and to sever that connection in order to make the interpretational theory conceptually simple and mathematically tractable, on the other hand. However, it seems to me that this is precisely what Davidson does in his Unified Theory of Interpretation.

António Zilhão
Universidade de Lisboa
e-mail: antonio.zilhao@mail.doc.fl.ul.pt



Philosophy of Mind

Filosofía de la Mente





IRRACIONALIDAD MOTIVADA: El caso del autoengaño

MONTSERRAT BORDES SOLANAS
Universidad Pompeu Fabra

ABSTRACT: This paper is an attempt to inquire into the nature of self-deception. I will afford a theory which links self-deception to wishful thinking and diverges from Mele (1997)'s motivational analysis of self-deception in that it (i) establishes as a necessary condition for self-deception the existence of a causal link between a desire and a belief unacknowledged by the self-deceived, unaware too of the counterevidential nature of his belief (focussed inferential blindness' thesis), (ii) allows only 'weak self-deception' cases and offers methodological reasons against the seemingly intentional and dissociative nature of self-deception, (iii) stresses that self-deception, unlike deception, is involved with epistemic reliability and not with truth value, and (iv) is able to distinguish between self-deception and mistake due to ignorance.

KEYWORDS: Self-deception, irrationality, epistemology, subpersonal mental level, counter-evidential belief-formation, akrasia, cognitive dissonance, motivation, wishful thinking.

En *De profundis* Oscar Wilde se compadecía católicamente de la miserable naturaleza de su amante, cuya mezquina conducta determinó los amargos años de prisión del escritor irlandés. A pesar de la lucidez con que describe su comportamiento egoísta hacia él, y dado el carácter disposicional del tipo de emoción implicada, Wilde exhibe un estado de autoengaño al afirmar sinceramente que Mr Douglas lo amaba.

En esta comunicación es mi propósito ofrecer un análisis conceptual del estado mental del autoengaño. Presentaré los dos modelos básicos de análisis rivales y, tras exponer las deficiencias que presenta el modelo interpersonal, defenderé una teoría que emparenta el autoengaño con la creencia desiderativa. Tras exponer el modelo interpersonal como un análisis que caracteriza el autoengaño como intencional y dissociativo (autoengaño 'fuerte'), presentaré las diferentes estrategias pensadas para eludir la atribución de inconsistencia al autoengañado en un esfuerzo por sintonizar con el principio de caridad. Argumentaré contra la estrategia más habitual, la compartimental, y presentaré el modelo no interpersonal de análisis en su versión motivacional como mejor

opción. Desde ese modelo, expondré una teoría más respetuosa con el principio de caridad y con el principio de exclusividad de la creencia, al no atribuir creencias conflictivas al autoengaño, siguiendo en parte el análisis de Mele (1997), pero que, a diferencia del mismo, (i) centra la clave del autoengaño en la existencia de un nexo causal deseo-creencia y en la incapacidad del sujeto para reconocer la inadecuación de su creencia con la evidencia (tesis de la ceguera inferencial focalizada), (ii) no acepta casos de autoengaño fuerte, argumentando desde premisas metodológicas contra su supuesto carácter intencional y disociativo, (iii) sostiene que el autoengaño tiene que ver con la falta de fiabilidad epistémica y no con la de valor veritativo y (iv) fija una diferencia radical entre el autoengaño y el mero error por ignorancia en cuanto que en el primer caso, y no en el segundo, se conoce la evidencia contraria a la creencia. Entenderé el autoengaño como un estado doxástico autocéntrico, irracional desde el punto de vista epistémico por su base motivacional pero diferente de la creencia desiderativa por su relación con la evidencia relevante.

El autoengaño es un ejemplar de lo que la tradición filosófica ha dado en calificar como pensamiento irracional y cuyo interés reciente renace en la medida en que se advierte que entre las actividades mentales de las personas de carne y hueso pocas son ejemplificaciones circunstanciales de racionalidad pura, muchas de racionalidad imperfecta y otras de simple irracionalidad. Junto con la acrasia, la creencia desiderativa y los casos de preferencias adaptativas como el fenómeno denominado de "uvas verdes", el autoengaño es el resultado de un proceso epistemológico ilegítimo en la formación y retención de creencias y/o decisiones, un caso de disonancia cognoscitiva, en la terminología de Festinger (1957). El autoengaño, como los otros fenómenos aludidos, es también irracional. La irracionalidad a la que me refiero es irracionalidad *epistémica*. En el caso del autoengaño la adquisición de la creencia es epistémicamente deficiente en su justificación al atentar contra un principio de racionalidad como el que enuncia Price (1965, 131): el grado de asentimiento a una proposición debe ser proporcional a la fuerza de la evidencia conocida acerca de esa proposición. El autoengañado, sin embargo, asiente a una proposición en contra de la evidencia disponible. En el caso de la acrasia, la irracionalidad se debe a que el individuo actúa contra su mejor juicio. Sin embargo, hay otro sentido de racionalidad según el que un estado o acción es racional en la medida en que es instrumentalmente eficiente en la obtención de un fin deseado¹. Y no se puede negar que muchos casos de acrasia y de autoengaño poseen un alto valor instrumental como mecanismos adaptativos, optimizadores del deseo general de conservar el equilibrio mental y reducir la

¹ Vid. los ensayos de Davidson (1980).

ansiedad. El autoengaño del oncólogo aprensivo sobre su propio cáncer² es benéfico no sólo para su salud mental, sino también para su salud física, dado que el reconocimiento de su enfermedad hubiera quebrado su disposición curativa.

Ensayistas como La Rochefoucauld, Montaigne o Pascal, tan interesados en el estudio de la etiología y relaciones funcionales de los sucesos de la mente humana, alertaban con frecuencia en sus textos acerca del uso generalizado de las estrategias de racionalización de nuestros actos, que ocultan los verdaderos motivos que los inspiran (el médico que cree haber escogido su profesión por vocación altruista y no, como realmente es el caso, por la reputación social que comporta). Algunos autores - desde luego, literatos - han llegado incluso a defender la universalidad del autoengaño en toda acción humana relevante. En el polo opuesto, los escépticos niegan su ocurrencia real considerando el autoengaño como un estado mental lógicamente incoherente y, en parte por ello, psicológicamente imposible³. En este trabajo me limitaré a examinar las teorías moderadas, ni universalistas ni escépticas, que aceptan la realidad circunstancial del autoengaño. Las principales teorías moderadas encajan en uno de estos dos modelos explicativos: interpersonal o motivacional no interpersonal. Quienes abogan por el modelo interpersonal analizan el autoengaño como un caso particular del engaño⁴. Para ellos 'X se autoengaña' ha de leerse según el esquema 'X engaña a Y' (engaño interpersonal), de modo que el autoengaño no es más que la versión reflexiva del engaño: 'X engaña a X' (engaño intrapersonal). Las condiciones necesarias y suficientes para el engaño pueden especificarse como sigue. Dados dos seres racionales, X e Y, y una proposición, p, X engaña a Y en un tiempo t' ($t > t'$) sobre que p si y sólo si:

(1) en t X intenta inducir en Y la creencia de que p.

(2) en t' X consigue llevar a cabo (1).

(3) en t y t' X sabe (o tiene la creencia verdadera de) que no-p.

Así, el engaño es un acto deliberado o intencional (por (1)), supone éxito en el intento (por (2)) y su objeto es una proposición verdadera (por (3)). En (3) se da cuenta de que no es necesario que X disponga de una justificación correcta o que haya adquirido su creencia en que no-p mediante un proceso causal fiable para poder engañar a Y. Basta con que no-p sea verdadera y que X así lo crea.

Según este modelo de análisis, el autoengaño ha de poseer las siguientes características: (i) ser deliberado o intencional (tesis de la intencionalidad del autoengaño), fruto de desear creer que p, y (ii) ser paradójico, ya que consiste en la posesión de creencias - o, al menos, actitudes proposicionales -

² El ejemplo es de Rorty (1972, 401).

³ Haight (1980).

⁴ Partidarios de este modelo son Demos (1960), Fingarette (1969), Rorty (1972), Pears (1975), Audi (1982), Davidson (1982) y (1985).

incompatibles. De (1)-(3) se sigue que en t' X cree que p y que $\text{no-}p$ (tesis de las creencias conflictivas). La tesis intencional afirma el carácter intencional o deliberado del autoengaño y la tesis de las creencias conflictivas su carácter disociativo. Llamaré 'autoengaño fuerte' al autoengaño que presenta un carácter intencional y disociativo a la vez. Nótese que si el autoengaño es intencional y X es conocedor del valor racional de la evidencia, entonces es disociativo, pero no a la inversa. Si X consigue autoengañarse al desear creer que p aun conociendo la evidencia a favor de que $\text{no-}p$, entonces X cree que $\text{no-}p$ (al menos al concebir el plan) y que p . Ahora bien, X puede tener esas creencias conflictivas sin que una de ellas haya sido intencional sino involuntariamente formada. Según este modelo interpersonal, tres condiciones necesarias han de cumplirse para que X se autoengañe acerca de que p :

(4) X cree que $\text{no-}p$

(5) X cree que p

(6) X cree que p porque X desea creer que p

(4) y (5) definen el carácter disociativo del autoengaño y (6) su carácter deliberado o intencional. Los partidarios de este modelo suelen apelar a la conducta elusiva de la evidencia que suelen presentar algunos sujetos autoengañados para exigir (4) y (6) (el marido que parece evitar toda confrontación con las pistas de la infidelidad de su esposa). Parece existir una *tensión* entre la tendencia a creer que p (verbalmente manifestada) coexistiendo con el conocimiento de que la evidencia va a favor de que $\text{no-}p$ (no verbalmente manifestada). Dada la contundencia de la evidencia contra la creencia acorde con la conducta verbal del sujeto y su ostensiva elusión de la misma, se juzga inviable que no la advierta. Como el sujeto elude constantemente la confrontación con la evidencia, se afirma que, aunque sabe que $\text{no-}p$, de algún modo ha conseguido manipular su creencia para llegar a creer que p . Es sintomático de los estados de autoengaño el que sus agentes malinterpreten la evidencia disponible y que exista un motivo por el que se produzca esa mala interpretación. Pero entonces, si el sujeto percibe la evidencia como amenazante, es que sabe que la creencia que sostiene tozudamente es falsa y toma medidas para eludir su poder persuasivo⁵. Éste es el argumento de la ceguera perversa, cuya conclusión es que el autoengaño es un tipo de ceguera voluntaria ante la evidencia.

Más adelante me referiré a la cuestión de la intencionalidad, pero ahora prefiero enfocar la característica notoriamente más problemática del autoengaño: su aparente inconsistencia (o su carácter disociativo sincrónico)⁶.

⁵ Vid. Davidson (1985).

⁶ Al caso diacrónico (en el que X cree que $\text{no-}p$ en t y p en t'), no paradójico, me referiré *infra*. Realmente, sólo el caso sincrónico responde exactamente al modelo de engaño

La lógica impide que las creencias conflictivas sean verdaderas a la vez, pero sólo el davidsoniano principio de caridad nos impide adscribir las a un mismo ser racional. Según este principio psicológico normativo a todo ser racional se le han de asignar creencias y deseos que optimicen su racionalidad. Y atenta contra este principio atribuir a un sujeto la violación de lo que se puede llamar la 'ley de exclusividad de la creencia': cuando se cree una proposición, se cree a expensas de sus rivales, esto es, creer que *p* implica rechazar las alternativas a que *p*. Puesto que el principio de caridad aconseja no atribuir creencias contradictorias conscientes a las personas - al menos cuando la incompatibilidad es obvia - los partidarios del modelo interpersonal han de buscar un modo de armonizar el principio de caridad con la atribución de autoengaño a un sujeto.

Los partidarios del modelo interpersonal suelen adoptar una de las siguientes estrategias elusivas de la atribución de inconsistencia al sujeto: (a) que las creencias conflictivas implicadas están parcial o totalmente aisladas entre sí. En la versión débil de esta estrategia, una de las creencias es inconsciente (estrategia de la creencia inconsciente o no advertida). En su versión fuerte, cada creencia está modularmente separada de la otra en compartimentos o niveles subpersonales distintos (inconscientes o no) de modo que están inferencialmente encapsuladas total o parcialmente (estrategia de la compartimentación)⁷. (b) que la actitud proposicional con respecto a una de las proposiciones no es la misma que con respecto a la otra (estrategia compatibilista)⁸. (c) que una de las creencias se sostiene en un momento de tiempo diferente (estrategia de la restricción temporal)⁹. (d) que las creencias en cuestión no se sostienen plena sino sólo parcialmente (estrategia de las creencias parciales)¹⁰.

La estrategia (a) ha sido, con mucho, la respuesta más habitual al problema que genera la tesis de las creencias conflictivas (lo que Mele (1997, 92) denomina el 'enredo estático' del autoengaño), a saber, la atribución de inconsistencia al sujeto, que cree que *p* y que no-*p*. Obviamente el problema no se da en el caso interpersonal a partir del que se construye el análisis del autoengaño. La doble instancia del engaño hace posible desde el punto de vista lógico la consistencia de la creencia de que *p* con la de que no-*p*, puesto que no es el mismo sujeto quien cree ambas cosas a la vez. Psicológicamente se explica

interpersonal, por cuanto que satisface la condición (3) según la que el engañador-engañado cree en *t'* que *p* y que no-*p*.

⁷ Demos (1960) sostiene la versión débil. La estrategia de la compartimentación la defienden Fingarette (1969), Davidson (1982) y (1985) y Pears (1982).

⁸ Haight (1980) dice que el autoengañado sabe que no-*p* pero cree que *p*.

⁹ Sorensen (1985).

¹⁰ Gibbins aboga por esta estrategia en sus comentarios a Mele (1997).

mi condición de engañado en la medida que la confianza depositada en el otro me ha impedido considerar toda la evidencia relevante para no dejarme engañar. Ahora bien, en el caso intrapersonal, si es sólo una la instancia implicada, si el mismo sujeto cree que *p* y que no-*p*, evalúa y no evalúa la evidencia relevante, la situación resulta absurda por inconsistente¹¹. Desde el modelo interpersonal esa inconsistencia es la base fundamental de la irracionalidad del autoengaño.

Se suele justificar la no autopercepción de la inconsistencia mental del sujeto postulando la existencia de creencias inconscientes. La inconsistencia es real, pero no es percibida por el sujeto. La atribución de inconsistencia no consciente (inconsciente o sólo preconsciente) a la que se suelen remitir las teorías de la compartimentación impide la violación flagrante del principio de caridad. Así, la madre autoengañada sobre la inocencia de su hijo creería inconscientemente que es un criminal, pero consciente y sinceramente insistiría en su honradez. Esta estrategia depende para su credibilidad de la definición de un criterio de identidad de estados mentales inconscientes¹². Ésta es una cuestión problemática, dada la peculiaridad de las condiciones de atribución o identificación de estos estados, que, por definición, no son inferibles a partir de criterios relativos a la conducta verbal. Si *X* afirma que cree que *p*, inferimos que cree que *p* y que es consciente de ello¹³. Ahora bien, como las creencias inconscientes son verbalmente inexpresables, los partidarios de la remisión a tales creencias apelan a la conducta no verbal¹⁴. Estos autores suelen pensar que el autoengaño presenta analogías notables con los casos de visión ciega o con el fenómeno phi con manchas de colores distintos y que el autoengaño nos da una razón más para descreer de una mente unitaria a favor de un modelo dennettiano de borradores múltiples. No obstante, no conozco ningún experimento relativo al autoengaño que no pueda ser interpretado sin atribución de una creencia inconsciente incompatible con la creencia verbalmente expresada. Cuando la conducta verbal sugiere fuertemente que el sujeto cree que no-*p*, siempre cabe la posibilidad de interpretar sus alegaciones a favor de que *p* como insinceras. Y es que la conducta no verbal no presenta la suficiente articulación como para constituir un criterio de identificación de un

¹¹ Aunque no voy a entrar en ello, porque mi objetivo es sólo conceptual-descriptivo, quiero señalar que sobre la paradoja epistemológica se superpone la paradoja ética: el creer que *p* contraevidencialmente puede inculpar al sujeto, mientras que el no creerlo lo exculparía. Piénsese, por ejemplo, en el caso de muchos alemanes autoengañados cuando veían desaparecer a sus vecinos judíos (el ejemplo es de Elster, 1979, 178). Ninguna de estas paradojas surge desde un análisis no interpersonal del autoengaño.

¹² Nótese que las creencias inconscientes no satisfacen la ley de exclusividad de la creencia. ¿Quizá hemos de considerarlas protocreencias?

¹³ "Una creencia lingüísticamente expresada es una creencia consciente" afirma Peacocke (1992, 154).

¹⁴ Sackeim y Gur (1979) describen e interpretan algunos experimentos al respecto.

estado de creencia: actuar *como si* se creyera que p no es actuar *expresando* que se cree que p. Por otra parte, no me ocuparé aquí de la cuestión epistemológica de las condiciones de identificación del autoengaño, sino tan sólo de sus condiciones de identidad.

En su versión fuerte, esta estrategia supone apelar a módulos subpersonales, secciones mentales que mantienen inferencialmente aisladas las creencias conflictivas y a otros estados mentales vinculados. Ahora bien, aunque esta estrategia sea recomendable quizá en el análisis de otros fenómenos mentales, no creo que sea preciso adoptarla en este caso. Según creo, presenta las siguientes deficiencias. Por un lado, parece negar la misma posibilidad del autoengaño: ya no es S el autoengañado, sino que subS1, que cree que no-p, inventa un proyecto para engañar a subS2, la auténtica entidad engañada, y que crea que p. Por otro, en cuanto a la ontología mental que comporta, postula una división permanente del yo con el fin de dar cuenta de un estado mental circunstancial. Desde luego, el autoengaño es un estado disposicional, no episódico (si fuera una enfermedad, sería más parecida a la tuberculosis que a la epilepsia), pero su duración en los casos no patológicos (no clínicamente esquizoides) hace que parezca excesiva y *ad hoc* la teoría de los compartimentos. Además, la compartimentación no resuelve el 'enredo dinámico'¹⁵ del autoengaño: cómo un sujeto adopta intencionalmente una creencia y se la apaña para ocultarse a sí mismo esa intención. Apelar a los homúnculos subS1 y subS2 tan sólo restablece la paradoja retrocediendo un paso, puesto que la postulación de entidades personales internas restablece las mismas incógnitas iniciales sobre su funcionamiento. La modularidad homuncular parece ser, además, la única respuesta disponible a este enredo desde el modelo explicativo interpersonal: si hay un proyecto intencional, el homúnculo que lo conciba ha de poseer toda la información, pero entonces, ¿cómo se oculta a sí mismo el proyecto?

A partir de estos comentarios, no es difícil colegir que mi teoría del autoengaño no ofrece un análisis compartimental del mismo. Tampoco será una teoría basada en la analogía con el engaño. Algunos teóricos, como Mele (1997), no rechazan como inválido *tout court* el modelo interpersonal, creyendo que efectivamente existen casos excepcionales de autoengaño fuerte, aunque no los consideren prototípicos¹⁶; otros, en cambio, entre los que yo misma me cuento, no ven necesario reconocer la existencia de casos de autoengaño 'débil' (ni disociativo ni intencional, sino motivacional: la creencia

¹⁵ La denominación es de Mele (1997, 92). *Vid.* Elster (1979, 77).

¹⁶ Para Mele el autoengaño en casos paradigmáticos no es ni intencional ni disociativo, sino que es un caso de motivación desviada emparentada con la creencia desiderativa.

de que p viene causada por el deseo de que p) y que defender el modelo interpersonal es empírica y metodológicamente inadecuado¹⁷.

El partidario del modelo no interpersonal invita a desoir las sirenas de la analogía léxica engaño-autoengaño. Según la teoría que defenderé, el autoengaño es un estado doxástico, con contenido proposicional (no hay autoengaño simple o directo/proposicional por analogía con el ver simple/ver proposicional): autoengañarse sobre que p es creer (en función de cierta etiología irracional) que p. El autoengaño es el estado final de un proceso. No es posible un autoengaño instantáneo. Por un lado, se necesita al menos un intervalo de tiempo para que el deseo cause la creencia contraevidencial. Por otro, el autoengaño como estado supone una cierta retención de la creencia contra la presión de la evidencia en contra. En cuanto estado, es disposicional, no ocurrente: su núcleo cognoscitivo no es un pensamiento, sino una creencia. Aunque el autoengaño siempre es autocéntrico (en el sentido que precisaré en la cláusula (vii) *infra*), no siempre es egocéntrico. El autoengaño egocéntrico trata sobre estados de cosas uno de cuyos constituyentes esenciales es el sujeto autoengañado (sus estados mentales - el inconfesable amor del sacerdote por un ser distinto de Dios - o no mentales - la irreconocida dependencia fisiológica del drogadicto respecto de la droga). El autoengaño no egocéntrico trata acerca de estados de cosas físicos o de los estados de mentales de otros individuos o en los que el sujeto autoengañado es sólo un constituyente accidental.

Según la teoría que defiendo, dado un ser racional, competente y autónomo inferencialmente (es decir, que no tiene mermadas sus capacidades inferenciales generales y que no se halla bajo la influencia de drogas, hipnosis,...), X, y una proposición, p, si X se autoengaña sobre que p de t a t', han de satisfacerse de t a t' las siguientes condiciones (generalmente) necesarias y juntamente suficientes:

(i) la evaluación racional de la evidencia disponible apunta a que no-p, y X conoce o sospecha la existencia de esa evidencia.

(ii) X no infiere que no-p a partir de la evidencia disponible por inferencia de su deseo de que p.

(iii) X desea que p.

(iv) X cree que p.

(v) X cree que p porque S desea que p.

(vi) X no sabe que (v).

(vii) la creencia de X sobre que p es una creencia autocéntrica.

¹⁷ Partidarios del modelo no interpersonal que admiten sólo casos de autoengaño débil son Siegler (1968), Szabados (1973) y Bach (1981). Entre ellos, Siegler (1968) no ofrece un análisis motivacional del mismo.

Esta teoría del autoengaño se basa no en el modelo interpersonal sino en el patrón de la creencia desiderativa (por (v)). Es una teoría deflacionista, así que no se compromete con arquitecturas teóricamente costosas como la freudiana ni con compartimentos subpersonales ni con misteriosas intenciones inconscientes. Niega el supuesto carácter intencional y disociativo del autoengaño, aceptando sólo casos de autoengaño débil o motivacional. Fundamentalmente se trata de un análisis que entiende el autoengaño como un caso de creencia contraevidencial motivada. La precisión en (i) que exige que X conozca o al menos sospeche la existencia de la contraevidencia es imprescindible para distinguir los casos de autoengaño de casos de mera ignorancia no motivada por falta de atención o examen superficial fruto del descuido. (i) señala (en la primera parte de la conjunción) el carácter contraevidencial del autoengaño y (en la segunda parte y junto con (v)) dibuja la diferencia entre error por ignorancia y autoengaño. En el segundo caso y no en el primero, la causa de la creencia se debe al papel causal de un deseo y el sujeto conoce la evidencia aunque no la reconoce o interpreta como apuntando contra su creencia. Así, la mujer del adúltero sabe que su marido regresa a casa más tarde de lo habitual, que su carácter es anormalmente hosco,...pero atribuye su conducta a otras causas menos preocupantes. Ello se debe a que la presión del deseo de que le sea fiel produce en ella una especie de ceguera inferencial que le impide extraer las consecuencias más dramáticas. Recordemos que los partidarios del modelo interpersonal aludían al carácter perverso de la ceguera del autoengañado. Ahora bien, no cabe atribuir tal perversión. Si el argumento de la ceguera perversa parecía convincente quizá se deba a que es simplemente circular. Desde luego, si la evidencia se elude *porque* se ve como una amenaza, entonces ha de reconocerse *como tal*. La cuestión, empero, es si en el autoengaño la evidencia se elude o simplemente no se conoce, o bien si se conoce pero no se extrae de ella las consecuencias racionalmente inferibles. Me inclino más bien por esta última opción. Según creo, la fuerza causal del deseo produce una especie de ceguera inferencial focalizada en el sujeto, de modo que, incluso con la evidencia ante sus ojos, el sujeto es incapaz de inferir lo que cualquier otro (o él mismo) en condiciones desapasionadas podría inferir. Así que el autoengañado se dirige directamente a la creencia que desea que sea verdad; no forma primero la creencia y luego la rechaza porque desea crearla falsa, como producto directo de un plan epistemológicamente perverso. Lo que desea el joven que presenta su artículo a una prestigiosa revista no es creer que le publican su artículo, sino verlo publicado. El autoengaño se ve así como un tipo de exposición selectiva a la información que comporta una disfunción inferencial temporalmente circunscrita a la situación relevante.

La ceguera inferencial focalizada del autoengañado es doble. Por un lado, es ceguera introspectiva, ya que el sujeto desconoce el nexo causal que vincula su deseo de que p con su creencia de que p. Se puede decir, que ese nexo es

temporalmente inaccesible a la conciencia del autoengañado, es decir, inaccesible a la maquinaria del razonamiento y control de la acción. Ese desconocimiento cuenta como contraejemplo prototípico de la tesis cartesiana de la transparencia de los estados mentales para su poseedor y de la incorregibilidad de la introspección. De hecho, en los casos de autoengaño es habitual que los testigos externos formulen diagnósticos mejores al respecto. Por otro, es ceguera extrospectiva, por ser incapaz de realizar la inferencia que va de la evidencia que conoce o sospecha a la creencia no deseada. Pero esa ceguera inferencial no es, en absoluto, ceguera voluntaria, efecto de una supuesta decisión de creer contraevidencialmente. Quizá una analogía desde la filosofía de la ciencia sea pertinente aquí. El deseo de que *p* parece ejercer en el autoengañado la función del cinturón protector lakatosiano que protege a la teoría científica de su revisión en función de la posible evidencia contraejemplar. El papel del autoengañado sería como el del científico defensor de la teoría del flogisto. En la explicación de la combustión de los metales y ante la evidencia de que el producto final de la combustión tiene un peso mayor que el del metal de origen no concluiría que la combustión no puede suponer una fuga de flogisto, sino que insistiría en defender su peso negativo. Las creencias del autoengañado son conservadoras hipótesis *ad hoc*, basadas en modificar los supuestos auxiliares sin justificación válida. A diferencia de la hipótesis *ad hoc* sobre la existencia de Neptuno, la evidencia no es compatible con la creencia de origen, sino que nos obliga a modificarla en su carácter de anomalía.

Algo distintivo del autoengaño es que la adquisición y persistencia de la creencia es tozudamente opuesta a la evidencia disponible, pero no que la creencia adquirida sea falsa. Un requisito de ese tipo parecería más bien un resabio injustificado de la falsa analogía engaño-autoengaño. Autoengañarse sobre que *p*, como tener la creencia desiderativa de que *p*, es una cuestión relativa a lo que puede o no ser racionalmente inferido a partir de la evidencia disponible sobre que *p*, independientemente de la verdad o falsedad de '*p*'. Es una cuestión de fiabilidad epistémica, no de valor veritativo. Después de todo, el hijo de la madre autoengañada podría ser inocente, a pesar de la extraordinaria evidencia en su contra. La diferencia lógica entre las condiciones de verificación de un enunciado y sus condiciones de verdad sirve como argumento para evitar inferir la falsedad de '*p*' a partir de la satisfacción de la condición (i). Una observación contrafáctica puede sernos útil para advertirlo más claramente: si ninguna otra evidencia extra apoyara la creencia de la madre excepto su deseo de que su hijo sea inocente, la verdad ocasional de su creencia no nos proporcionaría razones para atribuirle un estado diferente del autoengaño.

Como justificación de (vii) diré que nadie se autoengaña sobre una proposición que no ocupa ningún lugar en su espacio emocional-evaluativo. Si

me es indiferente que p o que no-p, si no me concierne personalmente su verdad o falsedad, entonces p no es un objeto candidato a un posible estado mío de autoengaño. En esa implicación personal (conocida o no por el sujeto) consiste lo que denomino el carácter 'autocéntrico' de la creencia del autoengañado.

Por último, mi respuesta al problema dinámico es su disolución: el autoengaño no es intencionado, aunque sí motivacional. La cláusula (v) establece que el estado de autoengaño se debe a la existencia de un nexo causal entre el deseo de que p y la creencia de que p. La mera existencia del deseo no es suficiente para que se produzca autoengaño, ni reconocer la existencia del deseo basta para superarlo, sino que es preciso reconocer que la única causa (una causa que no es una razón) de la creencia era el deseo: eso es suficiente para que un ser racional rechace su estado al valorarlo como epistémicamente deficiente. La mera presencia del deseo no determina la adquisición de la creencia relevante. Tampoco el reconocimiento de la existencia del deseo contraevidencial no es ni condición necesaria ni suficiente para que se mantenga o se dé un caso de autoengaño. No es suficiente: aunque la madre autoengañada puede desear que su hijo sea inocente, puede ser lo suficientemente lúcida para ver la cruda verdad, de modo que su deseo no interfiera con su evaluación racional de la evidencia. No es necesario: aunque la madre sepa que lo desea, puede estar autoengañada, en la medida en que no advierta que su deseo es la única causa de su creencia en la inocencia de su hijo. Lo que crucialmente ignora el autoengañado y que constituye la base de su deficiente estado epistémico no es el contenido de un estado, sino el que se dé una relación no razonable entre un deseo y su creencia (la cláusula (vi) recoge esta condición). Muchos teóricos del autoengaño suelen afirmar que el autoengañado reconocería su estado como deficiente en ausencia de ese deseo. Sin embargo, en mi enfoque, no es su mera presencia, sino su papel causal lo que realmente mantiene y genera el autoengaño. La diferencia que justifica los distintos estados de X e Y radica en que en X el deseo, por su diferente papel funcional en su red de estados mentales, no ejerce el poder causal que ejerce en Y.

BIBLIOGRAFÍA

- Bach, K. (1981): "An analysis of self-deception", *Philosophy and Phenomenological Research*, 49, 351-370.
- Canfield, J.V. y Gustavson, D.F. (1962): "Self-deception", *Analysis*, 23, 32-36.
- Davidson, D. (1980): *Essays on actions and events*, Oxford, Clarendon Press.
- (1982): "Paradoxes of irrationality", Wollheim y Hopkins, eds. (1982, 289-305).
- (1985): "Deception and division", Lepore, E. and McLaughlin, B. (1985, 138-148): *Actions and Events*, Oxford, Blackwell Publ.
- Demos, R. (1960): "Lying to oneself", *Journal of Philosophy*, 57, 18, 588-595.
- Elster, J. (1979): *Ulysses and the sirens*, Cambridge, Cambridge University Press.
- (1999): *Alchemies of the mind*, Cambridge, Cambridge University Press.
- Festinger, (1957): *A theory of cognitive dissonance*, Stanford University Press.
- Fingarette, H. (1969): *Self-deception*, Londres, Routledge and Kegan Paul.

- (1982): "Self-deception and the 'splitting of the ego'", en Wollheim, R. y Hopkins, J., eds. (1982, 212-227): *Philosophical Essays on Freud*, Cambridge, Cambridge University Press.
- Haight, M.R. (1980): *A study of self-deception*, Sussex, The Harvester Press.
- Mele, A.R. (1983): "Self-deception", *The Philosophical Quarterly*, 33, 133, 365-77.
- (1987): *Irrationality. An essay on akrasia, self-deception and self-control*, Oxford, Oxford University Press.
- (1997): "Real self-deception" y comentarios+réplicas, *Behavioral and brain sciences*, 20, 91-136.
- Peacocke, C. (1992): "The concept of belief: self-knowledge and referential coherence", en Peacocke, C. (1992, 147-176): *A study of concepts*, Cambridge MA: The MIT Press.
- Pears, D. (1975): "Paradoxes of self-deception", Pears (1975, 80-96): *Questions in the philosophy of mind*, Londres, Duckworth.
- (1982): "Motivated irrationality, Freudian theory and cognitive dissonance", Wollheim (1982, 264-288).
- (1984): *Motivated irrationality*, Oxford, Oxford University Press.
- (1991): "Self-deceptive belief-formation", *Synthese*, 89, 393-405.
- Sackeim, H.A. and Gur, R.C. (1979): "Self-deception: a concept in search of a phenomenon", *Journal of personality and social psychology*, 37, 2, 147-169.
- Siegler, F.A. (1962): "Demos on lying to oneself", *Journal of Philosophy*, 59, 469-475.
- Sorensen, R.A. (1985): "Self-deception and scattered events", *Mind*, 94, 64-69.
- Szabados, B. (1973): "Wishful thinking and self-deception", *Analysis*, 201-205.

Montserrat Bordes Solanas
Universitat Pompeu Fabra
e-mail: montserrat.bordes@humana.upf.es

PERCEPTUAL CONSCIOUSNESS AND THE REFLEXIVE CHARACTER OF ATTENTION*

OLGA FERNÁNDEZ PRAT
Universitat Autònoma de Barcelona

ABSTRACT: I suggest an account of perceptual judgements with demonstrative content, which is congenial to Searle's Fregean proposal of postulating a token-reflexive approach to the contents of perceptual beliefs. In the case of demonstrative thoughts, certain ways of "pointing out" to an object must be somehow included in the content, and for this end I substitute acts of attention for Searle's visual experiences. Moreover, I will roughly show how this turns out to be an externalistic account, and I point to the role of phenomenal properties in this account.

KEYWORDS: Attention, externalism, perceptual belief, perceptual demonstrative, Searle.

I

In this paper I will try to show how certain philosophical difficulties in the account of the intentional content of perceptual states can be overcome with the help of some conceptual developments which take as their point of departure some psychological notions and hypothesis.

According to John Searle's account in his book *Intentionality*, when I am perceiving a tree in front of me, I am thereby having a visual experience, which involves my consciously experiencing certain phenomenal properties; further, this visual experience represents a state of affairs consisting in the presence of a tree in front of me, and represents this state as causing the visual experience itself; finally, this is indeed the case, that is, to say, the state of affairs constituted by the state consisting in the presence of a tree causing the visual experience obtains.

* Part of this paper was read at the Eighth Annual Conference of the European Society for Philosophy and Psychology, University of Warwick, 23-27 July, 1999. I thank Daniel Quesada for encouragement and helpful comments. Financial support for this research was provided by the research projects PB97-0192 funded by the DGES, Spanish Ministry of Education, and HB97-0147 jointly funded by the DGES and the British Council.

In this way, Searle tries to answer the well known criticisms to (classical) Fregean accounts of content that point out their failure to secure the right particular-directedness of certain mental states to objects. Specifically, Searle has given an account in which a mental state is directed to a particular in a way that exploits the fact that the particular is currently in the subject's field of vision. Searle defends that the propositional content of perceptual experiences is essentially token-reflexive. The content of a visual experience is to be made explicit by "referring" to the very same particular visual experience enjoyed; he maintains that the same holds for the content of a corresponding perceptual belief. To repeat, the content of a visual experience or of a perceptual belief is specified, that is to say, it is made explicit, by "referring" to the same particular visual experience enjoyed. Searle's example, applied to the case of an experience is the following. The content of a perceptual belief of someone who dimly glimpses a man in the distance and that he could verbally express by saying "That man is wearing a red cap" can be expressed (in the right circumstances) in this way : "There is a man there causing this visual experience and that man is wearing a red cap" (p. 212). He tries in this manner to make compatible direct realism in perception with his professed internalism.

McDowell [McDowell, 1991] develops a convincing objection to Searle's theory from the complaint that Searle has not developed his theory enough. The fact that Searle is not concerned about the particular-directness signalled by "this visual experience" is, according to McDowell, a crucial deficiency in Searle's proposal and it marks and when developed there is a tension with his general internalist views of mental content.

I agree with the general direction of McDowell criticisms, which I will not review here. Nevertheless, I think that Searle's idea of the token-reflexivity of perceptual contents is a solid insight which a theory of perceptual content must maintain. My account tries to keep those insights without the undesirable consequences that Searle endorses.

First of all, I will try to suggest the guidelines of a token-reflexive account of demonstrative thoughts which exploits basically the same fact exploited in Searle's proposal, namely, the fact that the particular is currently in the subject's field of vision. Specifically, I am concerned with perceptual beliefs whose content one could express demonstratively as in, for instance, "That is dirty", "That carpet is torn" or "That is an oak". Such thoughts concern perceptually salient features of the subject's immediate surroundings. These are the thoughts that Evans described as being "information-based" in the sense that they are prompted by perceptual information from the object thought of [Evans 1982, p.145]. Just by looking around the room and focusing, say, on any object in it, I can entertain a

demonstrative thought (that is constituted by perceptual demonstratives) about that object.

II

Let us focus on the simplest cases in which demonstratives are used in a communicative situation, those in which a speaker is referring demonstratively to an object that is manifest to him and to the hearer as well. To be manifest means, here, that both speaker and hearer, who grasp the use of the demonstrative in that situation, could verify directly which object must be identified simultaneously with the uttering of the sentence containing the demonstrative, say 'this is a chair'. The referring intentions of the speaker have been expressed ostensively, e.g. by stretching out his arm and pointing to the object making the object referred perfectly evident to the hearer. Then, a successful demonstration has been made¹: the speaker's manifest deictic intentions correspond to the referring intentions of the speaker².

In order to develop a semantic theory of demonstratives the key analyzed elements in a situation of communication are demonstration and preference. How one defines these concepts and which role are made to play signals the differences between semantic theories.

Now, the fact of analyzing a situation of communication establishes guidelines that are not possible to follow when we try to make the move to a situation in thought. Of course, the different nature of the situations implies different constitutive elements. However, certain "translations" of the components from one theory to the other one are not only possible but useful. In my case, in order to make the "translations" an specific token-reflexive theory about *linguistic* demonstratives which I will review quickly now putting aside the arguments in defence of it [see, García-Carpintero, 1998].

Let us consider an utterance of sentences like (1), in circumstances in which a subject is pointing one after the other to branches of what we can consider different trees.

(1) That is an elm and that is a beech.

As García-Carpintero maintains, an adequate demonstrative theory must fulfill the following desiderata in relation to (1):

¹ That is, for the moment, I am leaving aside cases like the ones described by Kaplan where the speaker's manifested deictic intentions do not correspond to the referring intentions of the speaker.

² Kaplan calls to these intentions 'primary intentions'.

(i) The theory must give us something that is the same for the two different tokens of 'that', that is, something that corresponds to the meaning of the type-expression.

(ii) The theory must provide for the possibility that the two tokens of 'that' have different "contents" —different references—, since —in agreement with our linguistic intuitions— it is not to be presupposed that the two occurrences of the demonstrative-type must refer to the same thing.

The theory fulfills the two desiderata described above. The first one is satisfied by claiming that the use of demonstratives is subject to a linguistic rule roughly as follows: Any instance of 'that' refers to the individual (of the contextually determined sort) demonstrated when that instance is uttered.

In semantics we describe the referred object as the object referred in relation to a certain demonstration. How can we describe the way the object is determined in the context of (perceptual) thought? Clearly, in thought there should be an indication in the same way in which in simple cases of communication demonstratives must be accompanied by ostension, that signals singularity or individualizes the object to which the deictic intentions are directed. It seems to me that in order to have a demonstrative thought we pick out or select an object among all the objects in our immediate environment. In common language when we select in this way an object we say we are attending to it. There is a cognitive function that makes this possible. It is described by psychologists as the mechanism of attention. So, attention points to an object by making it manifest to the subject. A general way of describing demonstrations as the act that make manifest to the hearer what is the object of attention of the speaker. There are different methods for making salient an object in communication. Screaming, pointing to, winking at, all of them can be useful depending on the context. In thought the act of pointing to is realized by attention.

The other concept that is present in the analysis of the communication cases is 'utterance'. How can it be "translated" in thought? Let us remember the case in semantics:

1) Any instance of 'that' refers to the individual (of the contextually determined sort) demonstrated when that instance is uttered.

In perception I would describe the contribution of the demonstrative roughly in the following way:

1') 'That' refers to the object of attention in this act of attention.

In this formulation, 'acts of attention' replaces Searle's 'experiences'. This is because I maintain that the individuating character of acts of perception is better reflected in this fine-grained form. Furthermore, I will show how it fits well with the idea that perception is, in a certain sense, an intentional act and with the justification of token-reflexivity.

Now, Searle claims that the content of the perceptual belief of someone who dimly glimpses a man in the distance, which can ordinarily be expressed by saying, "That man is wearing a red cap" can be described (in the right circumstances) in this way:

2) There is a man there causing this visual experience and that man is wearing a red cap.

Correspondingly, I would describe the content of the mental alluded to, in which a subject is attending to an object in her field of vision in this manner:

2') There is an object presently attended to [of the class man] in this very act of attention and that object is wearing a red cap.

In both cases we can detect the presence of token-reflexivity. Like in Searle's formulation, where the visual experience itself is somehow included in the determination of the content, in my description the concrete act of attention itself, is included as a crucial individuating constituent. And, like in the parallel inclusion of the experience in Searle's account, this raises the doubt of whether the inclusion here of the act of attention in the content is not signalling a sort of indirection, in that the object referred is only "reached" by previously "capturing" an internal object, to wit, the concrete mental act of attention. But I think that we are here in a good position to explain why this is not so. So, I will explain now why this approach doesn't imply internalism, but it is instead externalistic, and I will defend it from the attacks of underdevelopment directed to Searle's own version by McDowell.

III

It is natural these days for a psychologist to think of perception in terms of the reception and procesing of information. Whatever happens at the level of the stimulation of nerve-endings must if perception —has thereby taken place— be such that action eventually performed as a result is to be characterized as performed in the light of information thus received. It is also natural to suppose that, if what is received is to function as information, it must be fitted into previously existing schemata, to use the term originally introduced in this context by F. C. Bartlett, and resuscitated more recently by, among others, Ulric Neisser [see Neisser, 1976]. Some philosophers have interpreted the idea of shemata in a strong way, in the sense that it implies that what is perceived will not be perceived as anything unless the perceiver has prior concepts, knowledge and beliefs. Interpreted in this way, the claim seems to be opposed to the thoughts of another eminent psychologist, James Gibson, who has claimed that in vision, for example, the ambient optical array provides enough information in itself to make it quite unnecessary to make reference to anything that goes on inside the perceiver.

I believe that Neisser's idea of schemata is less demanding and more in consonance with the most acceptable part of Gibson's claims, and in any case it is suitable for my purposes.

Neisser's notion of perceptual cycle will help us here. Neisser introduces the notion with the use of the following example:

If we happen to see someone smile, for example, there may be information to specify (a) the shapes of his teeth; (b) the changing positions of his lips; (c) the fact that he is carrying out a certain culturally-significant act; i.e., smiling; (d) something about his mood, which may be cheerful or sardonic or merely polite depending on the context in which the smile occurs. When we perceive his mood, we are not engaged in the same perceptual cycle as when we are attending to his lip movements. We develop a different (though perhaps overlapping) set of anticipations; we pick up information that extends over a different span of time (...).

[Neisser, 1976, pp. 21-22]

A perceptual cycle, in Neisser's sense, consists of three elements: an anticipatory schema, specifying what kind of information the perceiver aims, and expects, to acquire; exploratory activity guided by the schema, such as looking or listening; and finally the information picked up as a result of this activity, which in turn modifies the schema. Neisser's central claim is that what is perceived is partly determined by the schema governing the relevant perceptual cycle. We can distinguish two substantive points here. First, it is up to the perceiver, at least to some extent, to decide which schema to adopt, e.g. whether to concern herself with someone's teeth or mood. So the content of a perceptual experience will be in this way sensitive to the perceiver's interests. Second, according to Neisser, for the most part, having a schema is an acquired skill. Which information is picked up, in other words, will depend in part on which perceptual skills are being brought to bear.

Having Neisser's notion of a perceptual circle in mind facilitates paying due attention to the role of deliberate selective attention in determining the content of perceptual experiences. It is a mistake to think that this role can be reduced to the physical activities which, on the orthodox view, exhaust the subject's active contribution to her perceptual experience. Rather, we should think of such attention as a case of what C. Evans calls interrogative attention, or 'the attention of a probing intelligence in search of the answer to some question' (C. Evans, 1970, p. 100). Thus the content of a visual experience, for example, depends in part on which question the subject is trying to answer by visually attending to her environment.

Now, Searle maintains, as I have said above, that his claims about visual experiences are empirical ones. I think that in vision, as well as in

other perceptual modes, there are indeed empirical considerations which he does not take into account and that can help us in understanding and defending a specific account. In fact, I believe, the issue can be substantially clarified by appeal to Neisser's notion of a perceptual cycle. As Roessler has noticed [Roessler, 1999], appeal to Neisser's notion implies that visual perception can be understood in terms of "looking to see whether ...", which is itself an intentional activity. So, in one sense, someone deliberately looking to see whether *p* (i.e. trying to acquire visual information as to whether *p*) can be said to know what she is doing, just because this is the intention informing her activity. We should expect that the intention of acquiring visual information as to whether *p*, say, which, according to my account, informs the corresponding act of attention, must be subject to some kind of *monitoring*, aimed at establishing whether the intention is fulfilled [Roessler 1999, p. 59]. The reason for this is the following: if we accept that perception is an intentional activity and that, when perceiving, subjects are involved in perceptual cycles, one following the other, there must be in general clear connections between a perceptual cycle and the next. But this implies that subjects need to have information about the result of the cycle, and very specially information about whether the intentional purpose that informs a cycle has been accomplished or not. Achieving this information requires in turn that subjects have some kind of monitoring to find out about this.

Now, this monitoring of the intention—and so, the monitoring of the corresponding act of attention—is not to be confused with a further (reflexive) act of attention directed to the (perceptual) act of attention itself. It seems to me that it is precisely if such a further act were implied that we would have here a sort of "indirection" and the danger of internalism. But this is avoided in the line of explanation just suggested. Monitoring the act of attention is fully compatible with the claim that in perception our attention is directed to the external object.

IV

I have described the intentional content of visual experience, but I have not made any mention of its phenomenal character. In this last section, I would like to make some remarks on it.

In the account I have sketched 'acts of attention' take the place of Searle's 'visual experiences' in the description of the intentionality of perception. It seems to me that the shift from visual experiences to acts of attention in the description of content not only preserves the insights of a token-reflexive theory, whatever these may be in detail, but results in a reinforcement of it. Furthermore, there is a line of justification for that account from two related claims. The first one is akin to the idea that visual

perception has the mind-to-world direction of fit. Since Searle's shows that his theory agrees with the idea that a visual experience is not the object of a visual perception, and the features which specify the perception's intentional content are not in general literally features of the experience, the parallelism between the theory I have exposed and Searle's one, and my final suggestion about monitoring, which tries to make clear how it is possible in the modified account not to conclude that we see our experiences (or that we attend to visual experiences) in any way, shows already that in this modified theory the idea that we see things around us by attending to experiences (or in the classical terminology by "seeing our experiences") is not endorsed. About this point, I think that the description of the content in terms of acts of attention is very clarifying, and I will return below to this issue of attention in relation to experiences.

The second claim is about the role of phenomenal properties. I do think it can be regarded as another advantage for the modified theory. To begin with, what we have here is another example of the complementary research in philosophy and psychology. But let us pause to make some more detailed even if brief remarks about this second claim on phenomenal properties.

We begin then, with Searle's formulation:

When I see a car, or anything else for that matter, I have a certain sort of visual experience. In the visual perception of the car I don't see the visual experience, I see the car; but in seeing the car I have a visual experience, and the visual experience is an experience of the car, in a sense of "of" we will need to explain. It is important to emphasize that though the visual perception always has as a component a visual experience, it is not the visual experience that is seen, in any literal sense of "see", for if I close my eyes the visual experience ceases, but the car, the thing I see, does not cease. Furthermore, in general it makes no sense to ascribe to the visual experience the properties of the thing that the visual experience is of, the thing that I see. [Searle, 1983, p.37-38]

Thus, Searle claims that visual experiences are conscious mental events and that, consequently, they have phenomenal properties. Belief, imagination, attention, sensory experience, all have in common, among other things, that are forms of consciousness. Nevertheless, their intentionality is realized in different ways. In the case of sensory experience, imagination and attention, the concurrence of phenomenal properties is normally needed³,

³ Since the findings of Weiskrantz and others on "blindsight" it is not just logically conceivable that the stimulation of the retina of the eye might result in a given organism in

while in belief, it does not. Now, it is not possible to (consciously) attend to an object if this attention is not guided by phenomenal properties. So with regard to the cases I study, that is, perceptual demonstratives, entertaining or grasping them requires the help of those properties.

The world that we effortlessly —and usually accurately— attend to consists of complex objects that are characterized by their shapes, colours, movements, and other properties. In perceiving we are attending to objects. Psychology places exactly the necessity of properties like colour or shape with regard to attention in order to perceive the objects of the world. The notion of monitoring helps in this task. When we are perceiving a ball, for example, the object of our attention is the ball or the features of it. In spite of being disagreement within psychology about the subpersonal mechanisms underlying perceptual attention, there is the widely shared idea that at the personal level perceptual attention serves a function of selection of particular properties and relations of objects and events in the environment, and because of this it processes phenomenal properties. In cases of mental occurrences in which perceptual demonstratives are implied, phenomenologically we experience that we are attending to the objects of the world. It is most obvious that this is the case when we compare this experience with another we can achieve, that is, when attention is focused to our experiences. It seems clear that pretheoretically we commonly distinguish both situations: when we are attending to objects or to our experiences. But it is very important not to draw the wrong consequences from this fact, as it seems to happen in (Merleau-Ponty, 1962)⁴. Merleau-Ponty's comment is that, as one proceeds through the active affairs of life, one seldom concentrates one's attention on the flow of one's visual experiences, but rather on the things they are experiences of. From the alleged fact that focusing attention to something alters the character of this entity, he concludes that something like visual experiences do not exist and that they are only the result of adopting an "analytic attitude". But, putting aside the issue as to whether we must accept here that an alteration of the character of things takes place and, if so, in which way, what follows from my position is that we cannot put these cases on the same footing. Indeed, phenomenologically the similarity between the two cases is to be found in a moment of "concentration", so to speak. However, as I have said, according to the notion of attention I have sketched, it is not right to say, literally, that we attend *to* the experiences as we can say that we attend *to* (things in the) world. That would imply that visual experiences would be themselves

the same intentional content as some perceptual states without there being anything that constitutes the experience of seeing.

⁴ The summary that follows is reproduced from Searle (1983), p.44.

shaped or coloured and so on. So, we cannot say in any way that visual experiences are the *objects* of our attention. The only objects of attention are objects in the external world⁵.

REFERENCES

- Eilan, Naomi: "Perceptual Intentionality, Attention and Consciousness", in A. O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1998.
- Evans, C.O.: *The Subject of Consciousness*. London: George Allen & Unwin Ltd., 1970.
- Evans, Gareth: *The Varieties of Reference*. Oxford: Clarendon Press, 1982.
- García-Carpintero, Manuel: "Indexicals as Token-Reflexives". *Mind*, 1998, 107:427, p. 529-563.
- García-Carpintero, Manuel: "Searle on Perception", *Teorema*, 1999, vol. XVIII/1, p. 19-41.
- McDowell, John: "Intentionality De Re". In E. LePore & van Gulick (eds.), *John Searle and His Critics*. Oxford: Blackwell, 1991, p. 215-225.
- Merleau-Ponty, Maurice: *The Phenomenology of Perception*. London: Routledge & Kegan Paul, 1962.
- Neisser, Ulric: *Cognition and Reality*. New York: W.H. Freeman and Company, 1976.
- Peacocke, Christopher: "Conscious Attitudes, Attention and Self-Knowledge", in C. Wright, B. Smith & C. McDonald (eds.), *Knowing Our Own Minds*. Oxford: Clarendon Press, 1998.
- Roessler, Johannes: "Perception, Introspection and Attention", *European Journal of Philosophy*, 1999, 7:1, p. 47-64.
- Searle, John: *Intentionality*. Cambridge: Cambridge University Press, 1983.

Olga Fernández Prat

Universitat Autònoma de Barcelona

e-mail: ofernandezp@seneca.uab.es

⁵ Undoubtly, there is a phenomenon here of attention that concerns experiences and that we must elaborate theoretically. The notion of 'occupied attention' which is developed by Peacocke [Peacocke, (1998)] in order to explain the case in which we lend attention to our thoughts could perhaps be useful in order to explain as well the attention we offer to our experiences, although this suggestion is offered here only as a very tentative idea.

KANT ON THE SPONTANEITY OF THE MIND AND THE NATURE OF THE SELF

DIEGO LAWLER

Universidad Nacional de Quilmes

ABSTRACT: This work attempts to analyse the Kantian notion of the spontaneity of the mind and the correlative implications which can be derived from such a notion in respect of the nature of the self within and beyond the transcendental framework. Firstly, it presents the connection between apperception and spontaneity. Secondly, it considers the nature of the self within the transcendental framework. Finally, it discusses Kant's equivocal account of the 'I' that thinks and suggests to abandon the transcendental framework to credit a real cognitive self for the notion of spontaneity. To accomplish this aim it explores briefly Merleau-Ponty's position.

KEYWORDS: Mind, spontaneity, apperception, the 'I' that thinks, cognitive real self.

I

In an often quoted passage Kant states that through mere apperception we are conscious of certain faculties, namely, understanding and reason, '... the action of which cannot be ascribed to the receptivity of sensibility' (A547-B575:427)¹. Since the spontaneity of the understanding is, for Kant, always in an operative contrast with the receptivity of the senses, the clear implication is that apperception provides us with a consciousness of this spontaneity, and that this consciousness is an inseparable component of the conception of ourselves as cognitive subjects².

The more immediate consequence is that insofar as, for Kant, the activity of thinking designates the activity of judging and this in turn pertains

¹ All the quotations from the *First Critique* belong to *Critique of Pure Reason*, trans. N. Kemp Smith:1989.

² I am following Allison's suggestion (1983:273-74). He has claimed, quoting Kant's remark in B419:376, that the apperception thesis is not just a reference point or a transcendental principle but a 'real mode of self-consciousness' (1983:273), and this amounts to the fact that apperception yields a 'consciousness of the activity of thinking' (1983:274).

to the spontaneity of understanding -which takes place in the act of synthesis of given representations according to rules or concepts-, the apperception thesis means also a consciousness of this spontaneity³.

Following this perspective, it is my view that the apperception thesis could be interpreted as asserting that all human -finite- experience is unavoidably reflexive. It is reflexive because it is a means to awareness of oneself and of one's act of thinking. Of course, it has always been difficult to say precisely what this means, but in this interpretation Kant is regarded as suggesting at least the following.

On the one hand, that apperception yields awareness of 'what we are doing'. On the other hand, that it yields also awareness of ourselves as the subject or agent of that act⁴. And doing an act of thinking or, to put it in other words, making a judgement is what articulates such awareness. The former awareness, I believe, is addressed directly by Kant when he points to 'synthesis ... as an act, ... [that] is conscious to itself'(B153:166). The latter is also clearly indicated when he writes, '[A]ll the manifold of intuition has ... a necessary relation to the 'I think' in the same subject in which this manifold is found' (B132:153), and he adds, '[F]or the manifold representations, which are given in an intuition, would not be one and all *my* representations, if they did not all belong to one self-consciousness'(B133:153).

In what follow I wish to examine roughly these two points.

Broadly considered the first point brings out what seems to be the very thought implicit in the apperception condition, namely, that cognitive consciousness requires that consciousness be apperceptive; that is, in our making judgements, according to the interpretation that is being pursued here, we must be implicitly aware of our having judged the world to be such and such, and then of its possibly not being such an such⁵.

In order to better understand this argument, however, mention must be made of the apperception thesis in a less abstract way. Let me now revisit and redraw it in terms of judgement as the minimal form of such awareness to cast light upon it.

Once judgement is regarded as the activity of taking something as a such and such, it is clearly observed that, when we are engaged in cognitive activity, to make a judgement means, at least, to take an indeterminate

³ Recall here the equivalence between thinking, an act of spontaneity, and apperception, stated by Kant in B134.

⁴ See, for example, Kant's *Anthropology* (1794:§ 24), where he traces a distinction between apperception and inner sense.

⁵ To see this suggestion in a contemporary remark see McDowell:1994:26.

something x as an F or, in Aristotelian language, that 'S is P'⁶. So, according to the apperception condition and interpreting the activity of performing an act of thinking in terms of making a judgement, it follows that one cannot make a judgement without knowing what one is doing. Apperception, then, means an activity, inseparable from the judgmental activity, of knowing that one is knowing. As Patton puts it, '... such a judgement [that 'S is P'] is not only an awareness that S is P: it involves also, *dimly or clearly*, the judgement 'I think that S is P' (1936:404-05). [The emphasis is mine].

What matters here is that in all cases of judgement the mind must not only combine the items in single consciousness, but it must also be conscious of what it is doing. And this taking by the mind is a spontaneous, inherently self-conscious activity of the subject. Not only is something that the subject does, it is something that it does for itself and, therefore, self-consciously⁷.

On this interpretation, then, the particular mode of cognitive self-awareness described is what Kant refers to as the apperception thesis. And This activity is not another thing that one does when one judges. Indeed, it is an inseparable component of the first order activity itself. And one way of stressing this would be to suggest that this apperceptive feature of experience is adverbial; that is, when I perceive or think, I apperceptively perceive or think, and as Chisholm suggests (1976:49): it is 'used to qualify ... [the] undergoing'. Then judgement, and experience insofar as it involves judgement, are inherently reflexive, self-referential activities.

As regards the second point, apperception yields also awareness of the subject of experience and of its continuity in any conscious act. But being aware of oneself as the subject of one act of thinking on the one hand, and being aware of the identity of this subject with prior acts of thinking or intending on the other must not be confused with the reference to something in the place of the subject, whose existence or occurrence could be established by looking into oneself -introspection-, nor with awareness of oneself as an object of representation. Therefore, when one looks carefully at this point, one finds precisely the view that the use of the word 'I' in this context only indicates the subject of consciousness without characterising it (see Kant:B158n.:169). In effect, apperception yields the awareness of bringing to any conscious act a subject identical with the subject of prior conscious acts. As Kant notes, the consciousness of the identity of the self is given by a consciousness of the 'identity of the function whereby it synthetically combines it [the manifold] in

⁶ This latter manner of presentation is used by Patton 1936:404-05.

⁷ Pippin (1989:23) also makes this point.

one knowledge'(A108:136). Considerations like these can also be found in the second edition version of the second Paralogism (see Kant:B407:369)⁸.

These remarks, I believe, must be understood as stating the subject's possible awareness of its identity through its different states. So, putting things in this way, it certainly does not involve any claim about personal identity⁹. All Kant seems to be arguing is that a subject must be capable of being aware that it is the same subject through many representations, it need not be aware of anything specific about herself. Therefore what I want to remark is that Kant seems to be saying that, for example, in any act of thinking or remembering, we are aware of that we are thinking or imagining without knowing who we are, or even if we are mistaken about our identity. I can have a series of experiences today and I can have a series of experience tomorrow; although these last series require a formal unity, it might be the case that a change of persons takes place. So, it would be another person who is saying 'I'. Through this way of looking at the issue, apperception does not involve, certainly, attempts at self-identification. For Kant, then, the conclusions of the rationalist psychology must be rejected. In effect, the way 'I' appears to itself in its memories tells nothing about which earlier being it was.

In the next section I wish to analyse this Kantian I unveiled through apperception.

II

Kant's account of the 'I' of the 'I think' presents a very contemporary ring. It anticipates, in some way, Anscombe's attempt to undermine the common-sense view of 'I'; that is, the referential view and the indexical view. She writes, 'I' is neither a name nor another kind of expression whose logical role is to make a reference, *at all*'(1975:60). This also seems to be the case of Allison's account. For Allison 'I' designates only 'something in general', which is to say that it does not refer to anything at all'(1984:275-76). Despite that Allison regards apperception as a type of self-knowledge, he attempts to cut off this peculiar form of awareness from any reference to a cognitive real subject. He argues that it only refers to the transcendental subject, as the subject of such acts. So, insofar as the transcendental subject 'can be characterised only as "something in general = x" ' (1984:281), we have a reference to a logical subject instead of a reference to a cognitive real one. From

⁸ There Kant intends to block the move from the transcendental unity of apperception to the identity of the person. For Kant, this unity does not involve the subject in judgements of her own identity.

⁹ See on this point Cassam's analysis (1997:158-61) of Kant's logical identity of the 'I', and Kant:A363.

this perspective we can see operating a mechanism of abstraction from the real cognitive subject to the transcendental one. To conflate both means to be committed to a 'category mistake'(see Quassam:1997:19-21). In this way, Allison saves the unknowability of the real self, remains in a metaphysically neutral ground, and points to the fact of spontaneity in thought while, at the same time, takes apperception to be a mode of self-knowledge about a vacuous subject: a mere 'x'. Nevertheless, Allison himself seems to muddy the water. He points that through apperception we are conscious of the spontaneous activity of synthesis. But he also adds that, '[T]here can be no activity without an agent, and to acknowledge the existence of an activity is to acknowledge the existence of something that acts'(1984:280). At least this remark is against his idea of a vacuous transcendental subject.

But matters are far less straightforward than all of this suggests. Although Kant's view in the Third Paralogism rejects any attempt at self-identification, he speaks frequently in a way that suggests that the 'I' of the 'I think' seems to display a function of reference. The most impressive statement is this: 'It is obvious that in attaching 'I' to our thoughts we *designate* the subject of inherence only *transcendentally*, without knowing anything of it either by direct acquaintance or otherwise'(A355:337) [The emphasis is mine]. So, if any sense is to be made of this claim, what we need is an account which can make sense of this quotation. Allison's response seems to stop short of the fact.

In what follows, I shall present a plausible interpretation that could fit coherently both Kant's conclusion in the paralogism and his peculiar remark on reference, namely, 'transcendental designation'¹⁰.

One way to shed light on this issue is to apply at this point Shoemaker's concept, namely, self-reference without identification. According to Shoemaker: 'if a say "I feel pain" or "I see a canary" ... [the] use of the word 'I' as the subject of my statement is not due to my having identified as myself something of which I know, or believe, or wish to say, that the predicate of my statement applies to it' (1968:558). In this sense, the use of 'I' in 'I think' could be regarded as a self-reference without ascription of properties to this 'I'; that is, it is not the case that, according to this interpretation, an identification of the reference occurs via a description. This form of reference, which was called 'nonascriptive' by Brook (1994:74-7), applies to the fact that I can refer to myself as myself without supposing, in the act of referring, that I have any properties. The core of this idea is, I

¹⁰ I have borrowed some aspects of this interpretation from Brook:1994:chapter 4:sect.1.

believe, to disentangle two different logical functions which are here at work. The act of attaching many thoughts to one subject is different from the act of attributing properties to an object. This is so despite the formal pattern, -i.e., one-many, that appears to be implied. Hence, the manner in which the 'I' of the 'I thinks' accumulates or compiles thought is not the form in which an object compiles or accumulates properties. The 'I' which thinks is not a subject of predicates (see Rosenberg:1986:11-28).

From this perspective Kant's claims mean that 'in attaching 'I' to our thoughts', the awareness of ourselves that is produced does not contain anything about how we appear; that is, there is no identification of the referent via description. In this way is also preserved Kant's transcendental idealism. The self in itself cannot be an object of knowledge. Its nature remains unknowable (see Kant:B158:169).

The moral of this story is that what 'transcendental designation' makes evident is that there are certain properties which arise through acts of apperception, namely, our awareness of our spontaneous acts of thinking and of ourselves as cognitive subjects of such acts. Moreover, in each of our acts of thinking we are aware that we were the previous subject of our acts without both being committed to make any claim about any material continual through time or offering any kind of knowledge about our real selves. These properties are transcendental properties, and as such, are captured by acts of reflection.

This appears to be a sort of 'third interpretative position' about Kant. It is not the case of those interpretations that claim that Kant is interested in the structure of knowing, not in the knowing subject. Such perspectives emphasise the point that the normative dimension to questions of knowledge always prevents a psychologist reduction of such questions. The main argument directly asserts that the spontaneity of the subject must not be mistaken with any ability of an empirical being. As Ameriks notes, 'the mind ... is simply not the agent of spontaneity in Kant's sense, and in fact we have no ground for asserting there is an 'agent' here, or in any case that it is literally mental, literally a self'(1982:289). To make this mistake is to be influenced by 'an uncharitable understanding of the notion of "spontaneity"' (Ameriks:1982:288-9). Contrary to these interpretations, the 'third interpretative position' claims that although there is a mind that presents transcendental properties, we only have access to them through acts of transcendental apperception, without asserting anything else about such a mind. At least, this 'third interpretative position' makes room for the spontaneity of thinking, and does not reduce it, as Ameriks (1982:265-6) actually does, to a mere reflection of our limited and contingent human

perspective¹¹. To conceive of a spontaneous mind with transcendental properties allows to potentially think of spontaneity as one of its properties, but to consider its essential active character -spontaneity- as part of a merely human perspective seems to transform it into a mystery.

Nevertheless, it is neither the case of the other extreme position. This 'third interpretative position' does not claim, as for example Walker does (1978:131-34), that the subject of the spontaneous activity of synthesis cannot be the merely phenomenal subject; rather it must be the real self. This argument, which appears to conflate the transcendental condition of knowledge with properties of the self as it is in itself, notes Kant's famous remark (1989:B157:168) that one can still be conscious that, absolutely speaking, one exists, even if one has no noumenal knowledge of how one exists. From this admission follows that 'the floodgates are opened' (see Walker:1978:134) and the restraints of idealism are washed out. However, for what we have named the 'third interpretative position', the self, whose activity we are conscious of in apperception, is not phenomenal, nor is it possible to identify it with the noumenal self. This account only suggests that one is conscious both that it exists and of certain of its properties. The former grounded in the spontaneous activity of synthesis; the latter on the basis of acts of 'transcendental designation'. Nonetheless, we are left with the impression that the source of spontaneous activity of synthesis remains unexplained. The agency of such activity seems to be posited somewhere between the noumenal and the phenomenal realms, but never conceptualised and strictly localised. Furthermore, although this 'third interpretative position' seems to make sense of Kant's 'transcendental designation', it also stops short of the facts. It established the view that apperception yields awareness of one's cognitive acts and of oneself as spontaneous and self-legislating. This coheres clearly with Kant's remark when he notes in the *Prolegomena* that '[w]hen an appearance is given us, we are still quite free as to how we should judge the matter'(1977:290:34). But it stops short before registering that both forms of awareness might be pointing to the existence of a subject, a self, or mind engaged in the judgmental activity of knowledge.

Kant in the *Prolegomena* refers to apperception as implying 'the feeling of an existence' (1977:334n.:75) of a subject for such activity. Furthermore, he writes in the *First Critique* that '[M]an ... knows himself through pure apperception'(A547-B575:472). Indeed, Kant seems to have been convinced that although we can never, even in our own case, have any direct

¹¹ Ameriks' position is based on his own suggestion, namely, 'the species theory' (1982: 265-66).

knowledge of an immaterial 'soul', its permanence during life is evident since 'the thinking being (as man) is likewise an object of outer sense' (B415:373). In the first edition version of the second Paralogism he contends that 'the common expression that men think' (A360:340) can only be justified by the assumption that the real being perceivable by an outside observer is the outer appearance of what is, in itself, a simple thing which thinks. Hence, evaluating these claims one has the clear cut impression that Kant actually makes some assertions about this subject who is responsible for spontaneous acts of knowledge. Nevertheless, the far-reaching implications of those assertions are never integrated into or articulated in his relevant arguments. Kant's interpretation of such subject is in the end a purely formal one, stressing some of the logical differences between the attributes of a subject and the attributes of an object.

At this point one has the alternative to follow ontological interpreters, as for example Gottfried Martin (1961:Ch. V), and to grasp the comprehension of the knowing subject relying on the active practical subject. However, as Ameriks' historical analysis makes clear (1982:193-227), Kant actually discarded the idea that our freedom can give us theoretical knowledge of ourselves. So, from the alternatives within the transcendental framework the most sense of Kantian remarks one can make is to consider them as unclarities and lapses into inconsistency and to rule them out because of that. As it is well known, Kant quite often seems to let his own speculative impulse run away with his ideas.

An alternative way is to abandon the transcendental framework. By doing this one will not be suspended in mid-air. From the outset there is the strong intuition given by our common sense that anyone, to use Strawson's words, 'capable of formulating such a thought [i.e., 'I'-thoughts] will have mastered the ordinary practice of personal reference by the use of personal pronouns; and it is a rule of that practice that the first personal pronoun refers, on each occasion of its use, to whoever then uses it' (1994:210). So, while keeping in mind this intuition, what we need is an explanatory diagnosis of Kant's equivocal account of the 'I' of the 'I Think', and an argument that allow us to make clear how the spontaneity unveiled through apperception can fit into the natural world; that is, be incarnated in a body. All of this will be the content of the next section.

III

In this section I wish to suggest that Strawson's reconstruction of Kant could function as a clue in order to disentangle the ambiguous way in

which Kant presents his account of apperception. In order to both make the most sense of Strawson's position and complete the 'explanatory diagnosis' of the I of the 'I-Think', I shall analyse the arguments of John McDowell and Quassim Cassam, which are both based on Strawson's re-elaboration of Kant.

The re-drawing of the notion of experience by Strawson (1966:24-33) brings about an interpretation of apperception that sheds a new light on what apperception can provide: it gives at least the minimal ground to make room for the notion of an empirical self (see Strawson:1966:163). If he finds in Kant the view that 'the 'I think' of apperception represents also the tangential point of contact between the field of noumena and the world of appearances'(1966:173), we can find in his perspective the very possibility for anchoring an account that accommodates a real self for the Kantian I without dismissing the notion of spontaneity. Our interpretation of Strawson could function as follows: the subject has both a universal, law-discovering side, by virtue of which she can build up a picture of independent phenomenal reality in space and time, and also an element of selectivity within that reality, by virtue of both her position in such reality and her spontaneity. Indeed, her spontaneity carves out a specific route of appearances -both subjective and objective- within such reality, and what is essential at this point is the fact that the spontaneous subject must self-ascribe experiences as her own experiences; that is, as experiences that pertain to a real empirical self. However we can only give a full sense of such self-ascription if there are applicable criteria of identity, by means of which a given self can be located and given a place among other selves in the space-time world (see Strawson:1996:102-103). Therefore, the subjective unity of self-consciousness can only make full sense if it can be regarded as one experiential route through the objective world. Furthermore, we do not identify our own subjective route of experience and ascribe it to ourselves by making use of any criteria whatsoever. In fact, we can only give a full sense to such ascription if our selves are also given a definite bodily location in the world. This perspective actually allows us to go further than Kant and develop that which is comprised in apperception; that is, the implications of possible self-ascription of experiences. From this follows that is not sufficient to have the capacity to ascribe experiences to an 'abiding self' (Strawson:1966:103), if this self remains vacuous, i.e., transcendental. So, to conceive of oneself as an embodied self is fundamental in order to make sense of an account of self-ascription of experiences.¹²

¹² This idea, of course, involves the recognition of the concept of a Strawsonian person (1959:ch.3).

Having this framework in mind, let me consider McDowell's arguments. The line of his attack on Kant's correlation between self-consciousness and consciousness of the world is composed as follows. The temporal continuity of the subject, that according to Strawson's notion of experience, is the other side of the coin of perceptions as perceptions of enduring objects, 'shrinks', within Kant's position, 'to the continuity of a mere point of view, not apparently, a substantial continuant' (1994:100). This is so because when one credits the 'I' of the 'I think' with a persistent referent one remains within the 'flow of consciousness' (1994:103), and in order to avoid a Cartesian ego one is left only with the idea of a formal identical 'I' through time; i.e., the Kantian 'I'. However, this presupposition can be discarded. Once we embrace the possibility of locating such a persistent referent in a 'wider context', we begin to understand that the subject of experience which appears in Kant as 'an abstraction from the ordinary substantial persistence of the living subject of experience' (McDowell:1994:103), could be conceived of as a real self. McDowell's candidate for occupying this place is Locke's concept of person (see McDowell:1994:100).

This conceptual version relies on a neo-Kantian argument put forward by Strawson (1966:162-70). The core of the argument is that to conceive of oneself as a real self, in the sense of being a physical object, is a necessary condition of self-consciousness. The reasoning which gives support to this argument, according to Cassam (1997:122-3), is articulated as follows. In order to make and understand a judgement about something one must have the capacity to discriminate the object of one's judgement from all other things. Therefore, one must know what kind of thing one's judgement is about. From this follows that when one grasps an 'I-thought' -that is, self-ascribe an experience-, one must be able to be in a position to discriminate that to which one ascribes the experience from all other things, and then one must know what kind of thing one is. Consequently, if experiences are ascribed to a person, and a person is actually a physical object, the process of ascribing experiences requires knowledge of what one is; that is, the requirement is to conceive of oneself as a physical object.

This argument, which Cassam (1997:122) calls '*The Second Concept Version of the Identity Argument*', shows some debilitation. Basically, they can be out-line in two points. First, this type of argument, as we already see in the case of McDowell, relies on Locke's concept of person, but it does not do anything in order to justify such a concept. So, it is not clear that those who endorse such an argument are committed to accept such a concept. Secondly,

this argument involves a belief-directed transcendental argument, and as Cassam notes (1997:127-9) one can raise objections to it. In effect, this argument cannot discard conclusively the Cartesian approach. If we follow a Cartesian line of thought, we can preserve our capacity to make 'I-thoughts' without having to believe that persons are physical objects in the world. As Cassam puts the point: '... even if it is true that self-consciousness subjects or person are physical objects among physical objects, it is not a necessary condition of their being self-conscious that they *believe* that this is so'(1997:127).

In order to address these difficulties the belief component of the argument must be replaced by an 'intuitive awareness' (Cassam:1997:136) component. This procedure will give place to an intuition version of such an argument, which is, in fact, Cassam's position.

In what follows I shall consider first Cassam's account of Kantian formality. This will conclude our 'explanatory diagnosis' of the Kant's equivocal account of self-consciousness. Secondly, I shall present the intuition version argument, and finally I shall point to some explanatory dissatisfactions that we are left with.

Cassam's position raises a serious objection to the Kantian 'independent intelligible thesis'. This thesis asserts that one can grasp the identity of form of different 'I-think' thoughts without being conscious of the 'numerically identical subject of different 'I-think' instances as an object among others in the world'(Cassam:1997:162). Clearly this thesis is the basis which allows Kant to conceive of the transcendental I as purely formal. So, if it were possible to sustain this thesis against any particular objection, transcendental apperception would be fully intelligible without relying on the applicability of any empirical criterion of identity to the subject of experience. Cassam points out that this is not the case. The objection is as follows: if it shown that 'consciousness of the numerically identical subject of different 'I think' instances as a substantial object among others in the world is necessary for consciousness of such instances'(1997:164), then it follows that the 'independent intelligible thesis' is untenable.

The argument of the objection is based on the '*relativization requirement*'(1997:165), and is articulated in the following respect: in order to think the identity of different 'I think' thoughts, one must consider such thoughts as merely pure forms irrespective of the referent of the 'I'. But, given that in order to be able to specify the content of such types of thought, one must ascribe them to an individual person, then 'on the assumption that persons are object among others in the world' the Kantian formal account of

the I is not feasible. Therefore, the moral is that the 'impersonal' perspective of transcendental apperception does not pass the test (Cassam:1997:165).

Let me consider our second concern. The intuition version argument claims (Cassam:1997:136-53) that the manner in which one knows which thing one is does not have to involve the conception that the subject of one's thought is a physical object. By way of contrast such a manner has to be supported by an intuitive awareness of that to which one ascribes one's representations as 'shaped, solid, and located'(1997:141). This argument, at least, removes the objections against the concept version. In effect, it is by way of being intuitively aware of one's spatial properties that it is possible to accomplish the discrimination requirement on self-reference, which in the concept version remains incompletely achieved.

Despite the achievement of the intuition version argument, we are left with some explanatory dissatisfactions. Perhaps the most important is the lack of any account of what seems to be a basic insight that belongs from our common sense: the fact that in order to be aware of ourselves as 'shaped, solid and located', we must presuppose our own corporeality. If this is the case, what seems to be relevant is an account of the kind of awareness of oneself which is involved when one's own body, experienced as one's own body, is part of the field of one's experience. However, an analysis of such a type of experience not only remains beyond Cassam's intuition version, but also requires exploration of other philosophical frameworks. A support for this latter suggestion could be put as follows. Cassam's arguments are based on the effort for disarming the very obstacle for any account of the awareness of oneself as a real embodied subject: the Self-Consciousness Argument for the Exclusion Thesis (1997:9-21). This thesis denies that thinking selves are bodily presence in the world. One way to put this is to note, as such argument actually does, that what it is in a position of a subject cannot be an object for itself without being deprived of the very notion of the subject. So, if that is the very concern what we need is to attack directly this epistemological premise. The clue for doing this seems to find a basis for our awareness both of subject and object in some more fundamental experiential ground which will be at the same time both pre-objective and pre-subjective.

To summarise our result. If we now collect together the claims so far exposed, we have the following picture. Strawson's reconstruction notion of Kant's experience prompts the thought that spontaneity unveiled through apperception can be credited with a real self. Apperception, reinterpreted by Strawson, can make room for such an idea. In order to both make the most sense of this and to analyse the equivocal way Kant presents the account of

the 'I' of the 'I think' it was necessary to articulate McDowell's and Cassam's arguments. Following the latter we reach the final stage in which our desire for making sense of our awareness of our real bodily presence in the world shrinks to the point of dismissing the experience which includes our awareness of our own body as part of it. What we need now is to make sense of that kind of experience in order to anchor our incarnate presence as spontaneous subjects without, at the same time, losing the notion of spontaneity which we discovered in Kant's philosophy. What I would like to suggest is that Merleau-Ponty's position could fit this role.

Roughly, for Merleau-Ponty the subject actually engaged in the activities of perceiving and knowing the world, and the conditions without which such activities would fall apart into incoherence, are connected with the dialectic relation between our bodies and the things. By way of contrast, although Kant identified clearly the epistemic construalabilities in order to have knowledge of the world and the peculiarities of spontaneity of apperception which, conceived as belonging to the self as activity or an agent of synthesis, constitutes the clue to our epistemic capacities, he seemed to leave out the problem of the body of such agent or self. In fact, as we have already mentioned, this omission of the real self is the counterpart of the problematic way in which the 'things in themselves' are placed in his philosophical framework. For Merleau-Ponty, however, the very phenomenon of the world is articulated around the activity of our bodies, and this is prior to the beginning of any analytical reflection (see Merleau-Ponty:1962:40). He says: '...my experience breaks forth into things and transcends itself in them, because it always comes into being within the framework of a certain setting in relation to the world which is the definition of my body'(1962:303).

Nevertheless, the way in which the concept of spontaneity is articulated in Merleau-Ponty's concept of the body is not an easy task to elucidate. In order to situate the Kantian transcendental and spontaneous subject -which seems to be located outside us, everywhere- in a bodily presence, he begins with putting forward the question: 'who is thinking?' (1962:62) in each act of thought. Once this question is formulated, we can disclose the basic unit of his new notion of understanding: the body. He writes: 'The body is the vehicle of being in the world, and having a body is, for a living creature, to be involved in a definite environment, to identify oneself with certain projects and be continually committed to them' (1962:82). This is, in fact, what is revealed as the essential feature of our perceptual experiences. So, the Kantian subject must be grounded in a bodily presence. It constitutes a world, however, in order to 'be able to assert a true, the actual subject must in

the first place have a world or be in the world' (1962:129). To have already a world means, in this case, to be encompassed by a 'system of meanings' that do not 'require to be made explicit in order to be made exploited' (1962:129).

According to this account, the experience of our own body is the primary ground from which analytical reflection abstracts the concept of the subject and the one of the object, and '... gives us only the thought about the body, or the body as an idea, and not the experience of the body or the body in reality' (1962:198)

But matters are far less straightforward than all of this suggests. Indeed a complete account involves undertaking further work. Although I cannot pursue it accurately here, I shall close with a suggestion that I hope at least points to the arguments that must be developed.

For Merleau-Ponty, the synthesis performed by the body is, insofar as it is grounded in the action of a body-agent, a temporal synthesis. This discloses the subject of our primary perceptual experience as 'nothing but temporality' (1962:239). One of the consequences of this is that 'time is the foundation and measure of our spontaneity' (1962:428). This is the case, because, according to Merleau-Ponty, 'I am not the initiator of the process of temporalization; I did not choose to come into the world, yet once I am born, time flows through me, whatever I do' (1962:427). From this point of view, then, a complete account of spontaneity requires analysis of both the problems of the historicity of the subject and, with them, the respective notion of freedom. If this is the case, in order to elucidate completely Merleau-Ponty's position in respect to Kant, our analysis of Kant requires extension. The counterpart of Merleau-Ponty's notion of freedom is located, in Kant, in his practical philosophy. So, to complete our picture of Kant what is required is the analysis of the distinctions and similarities between Kant's notions of spontaneity of understanding and the autonomy of reason. Such notions appear to involve both indeterminacy by the world as it appears, by sensibility and experience, and governance by rational principles (see Kant:1989:A447-B475:410-11; A533-B561:464). But at the same time, these notions seem to mark the difference between the faculty of understanding and the faculty of reason. In this sense, the freedom required by morality seems to require a distinction from our spontaneity in knowledge.

Once this work has been done, the inter-relationship between these two philosophical traditions will be more comprehensible.

REFERENCES

- Allison, H., 1984, *Kant's Transcendental Idealism*, New Haven: Yale University Press.
 Ameriks, K., 1982, *Kant's Theory of Mind: An Analysis of the Paralogisms of Pure Reason*,

- Anscombe, G.E.M., 1975, "The First Person", in S.D. Guttenplan, ed., 1975, *Mind and Language*, Oxford: Oxford University Press, pp. 45-65. Bloomington and Indianapolis: Indiana University Press.
- Brook, A., 1994, *Kant and the Mind*, Cambridge: Cambridge University Press.
- Cassam, Q., 1997, *Self and World*, Oxford: Clarendon Press.
- Cassam, Q., ed., 1994, *Self-Knowledge*, Oxford: Oxford University Press.
- Chisholm, R., 1976, *Person and Object. A Metaphysical Study*, London: George Allen and Unwin Ltd.
- Kant, I., 1977, *Prolegomena to Any Future Metaphysics*, original trans. P. Carus, revised by J. Ellington, Indianapolis: Hackett Publishing Company.
- Kant, I., 1974, *Anthropology from a Pragmatic Point of View*, trans. Mary Gregor, The Hague: Martinus Nijhoff.
- Kant, I., 1929, *Critique of Pure Reason*, trans. Norman Kemp Smith as *Inmanuel Kant's Critique of Pure Reason*, London: Macmillan.
- Martin, G., 1961, *Kant's Metaphysics and Theory of Science*, Manchester: The Univeristy Press.
- McDowell, J., 1994, *Mind and World*, London-Cambridge, Mass.: Harvard University Press.
- Merleau-Ponty, M., 1964, *The Primacy of Perception*, ed. by James M. Edie, EEUU: Northwestern University Press.
- Merleau-Ponty, M., 1962, *Phenomenology of Perception*, trans. C. Smith, London: Routledge and Keegan Paul.
- Paton, H.J., 1936, *Kant's Metaphysics of Experience*, 2 vols., New York: Macmillan.
- Pippin, R.B., 1989, *Hegel's Idealism. The Satisfactions of Self-Consciousness*, Cambridge: Cambridge University Press.
- Rosenberg, J., 1986, *The Thinking Self*, Philadelphia: Temple University Press.
- Shoemaker, S., 1968, "Self-Reference and Self-Awareness", *Journal of Philosophy*, 65: 555-68.
- Strawson, P.F., 1994, "The First Person -And Others", in Q. Cassam, 1994, *Self-Knowledge*, Oxford: Oxford University Press: pp. 210-215.
- Strawson, P.F., 1974, "Imagination and Perception", in P. F. Strawson, 1974, *Freedom and Resentment and Other Essays*, London: Methuen.
- Strawson, P.F., 1966, *The Bounds of Sense, and Essay on Kant's Critique of Pure Reason*, London: Methuen.
- Walker, R.C.S., 1978, *Kant*, London-Boston:Routledge and Kegan Paul.

Diego Lawler

Universidad Nacional de Quilmes

He is currently doing his PhD. at Universidad de Salamanca

e-mail: dlawler@gugu.usal.es



CONTENIDO CONCEPTUAL Y NO CONCEPTUAL: ¿Autonomía o interdependencia?

FERNANDO MARTÍNEZ MANRIQUE
Rutgers University

ABSTRACT: This paper examines the distinction between conceptual and nonconceptual content. I contend that if the nonconceptual content is construed as autonomous, then either it cannot fulfill a foundational role with respect to conceptual content, or it becomes indistinguishable from it.

KEYWORDS: Concept, connectionism, non-conceptual content, mental representation.

Introducción

Las teorías de contenido no conceptual (Evans, 1982; Cussins, 1990; Peacocke, 1992; Bermúdez, 1998) son el penúltimo intento de ofrecer una visión de la vida mental que conecte el pensamiento regimentado conceptualmente con la riqueza del mundo experiencial de un sujeto pensante. De un modo general, (v.g. Cussins, 1990) se puede definir el contenido no conceptual (CNC) como aquel contenido que puede adscribirse a una criatura sin adscribirle por ello la posesión de los conceptos que se requieren para especificar ese contenido (dentro de una formulación canónica y relativa a una teoría, según Cussins, de las propiedades a las que hace mención dicho contenido). El afán del teórico del CNC es ofrecer una herramienta para fundamentar el contenido conceptual (CC). Más concretamente, el CNC tendría dos labores explicativas principales (Bermúdez, 1998, p. 58): 1) explicar en qué consiste poseer determinado concepto, 2) dar cuenta de la adquisición de habilidades conceptuales. Este intento fundacional se distingue de las teorías empiristas de los “datos de los sentidos” en que, a diferencia del carácter “dado” de estos últimos, el CNC es consecuencia de un proceso activo de la mente, que estructura la experiencia representacionalmente.

Si el CC se asienta sobre el CNC, cabe preguntarse cuál es el tipo de prioridad de la que goza el segundo con respecto al primero. Siguiendo a Davies (1998) podemos distinguir al menos tres tipos de prioridad: ontológica, epistemológica y analítica. Si atendemos a la formulación general del CNC expuesta más arriba, parece que el CNC no puede reclamar los dos últimos tipos de prioridad sobre el CC. De hecho, la impresión es más bien que existe una prioridad de lo conceptual con respecto a lo no conceptual tanto epistemológica, en cuanto que el conocimiento acerca del CNC procede a través del conocimiento acerca de los conceptos, como analítica, en cuanto que las nociones clave del estudio del CNC resultan venir analizadas en términos de nociones claves del CC. En otras palabras, dado que el CNC se formula de un modo negativo, como aquel contenido que escapa a su expresión conceptual, determinar si una criatura posee un determinado CNC supondría: 1) formular canónicamente la propiedad que constituye dicho CNC, 2) determinar si la criatura precisa poseer los conceptos que figuran en la descripción canónica. Claramente, este procedimiento implica una comprensión de lo que supone para la criatura poseer dichos conceptos, como requisito previo para determinar la posesión del CNC en cuestión. A lo sumo, el teórico del CNC sólo puede sostener, como Peacocke (1994) que CNC y CC (concretamente, el contenido de escenario y el contenido espacial) se elucidan simultáneamente.

A pesar de esto, el intento fundacional que alimenta las teorías del CNC lleva a afirmar la prioridad ontológica de éste sobre el CC. Es en este sentido en el que Bermúdez (1994, 1998) presenta su Principio de Autonomía (PA):

It is possible for a creature to be in states with nonconceptual content, even though that creature possesses no concepts at all.
(Bermúdez, 1998, p. 61)

Afirmar la posibilidad de CNC sin CC (por ejemplo, en animales sin lenguaje) no establece por sí solo la prioridad ontológica; para este fin es preciso complementar el PA añadiendo que no es posible que una criatura posea conceptos sin poseer (o haber poseído) CNC.

No todos los teóricos del CNC aceptan el PA. Peacocke (1994) expresamente lo rechaza¹. No obstante, no resulta posible establecer si el PA es correcto en ausencia de una caracterización más precisa del CNC, que nos

¹ El intercambio entre Bermúdez (1994) y Peacocke (1994) tiene en buena parte que ver con una lectura fuerte o débil del alcance del CNC. Peacocke sostiene una lectura débil en la que el CNC siempre requiere de la posesión de *algún* concepto, lo cual no entraña que sea totalmente captable por CC: para ello dichos conceptos deben dar cuenta exhaustivamente de la experiencia en cuestión. Bermúdez considera que aceptar esta lectura supone colapsar de hecho CC y CNC, por lo cual recomienda la lectura fuerte que conduce al PA.

permita fijar cuál es el contraste real con el CC. El CNC, al igual que el CC, se define como representacional, y la diferencia entre ambos radica en el tipo de representación (Bermúdez, 1998, p. 51). Pero mientras se dedica un gran esfuerzo a argumentar en favor de la representacionalidad del CNC, es menor el dedicado a aclarar en qué sentido la representación es diferente, aparte de en no requerir conceptos². El problema con el que se enfrenta el teórico del CNC, y que es el objeto de esta comunicación, es cómo caracterizar esta entidad de un modo más preciso de modo que evite, por una parte, hacerla tan débil que colapse en una noción asimilable a los “datos de los sentidos”, y por otra, hacerla tan fuerte que resulte indistinguible del CC, haciendo vacuo el intento de fundamentarlo.

Dado que el CNC se presenta de por oposición al CC, una posibilidad a la hora de determinar la diferencia entre ambos consiste en analizar las condiciones que algo debe cumplir para ser un concepto y determinar cuáles de dichas condiciones puede abandonarse en el caso del CNC. En esta comunicación voy a utilizar las condiciones que Fodor (1998) establece para una teoría de conceptos como punto de contraste. La lógica de este procedimiento establece que si no es posible para el teórico del CNC abandonar ninguna de las condiciones, entonces su noción de lo no conceptual es indistinguible de la noción de lo conceptual. No obstante, es posible objetar que si el CNC cumple todas las condiciones de esta teoría, no se sigue que sea indistinguible del CC, sino que es posible que sea la propia teoría de Fodor la que falle: quizás las condiciones determinen una teoría del contenido *en general* y no del conceptual en particular. Aun así, una ventaja de emplear la teoría de Fodor estriba en que es nítida y permite evaluar las consecuencias de abandonar una u otra condición.

Condiciones de Fodor para una teoría de conceptos

Fodor (1998, cap. 2) plantea las siguientes cinco condiciones que debe cumplir una teoría de conceptos, que funcionan a modo de restricciones respecto a la naturaleza de estos.

1. Los conceptos son particulares mentales.

² Un intento de ofrecer dos modos de constituirse una representación orientado a fundamentar la distinción entre la representación conceptual y la no conceptual es el de Dretske (1995), para quien podemos distinguir entre representación sistémica (Rs) y representación adquirida (Ra). Simplificando su posición, la Rs es fruto de la naturaleza del sistema en el que tiene lugar y sostendría el CNC; mientras que la Ra depende del desarrollo ontogenético del sistema en cuestión y sostiene el CC. No obstante, un problema con la Rs es que no está claro que cumpla la condición que para Dretske mismo distingue la representacionalidad, esto es, no está claro que permita la posibilidad de representación falsa (*misrepresentation*).

Un concepto se corresponde con un estado mental que puede ser causa y efecto de otros estados mentales. Más aún, un concepto tiene una existencia efectiva, en el sentido de que existe un suceso físico particular que instancia ese concepto y ningún otro, y en el sentido de que si dos conceptos son distintos entonces exigen dos sucesos físicos distintos que los instancien.

2. Los conceptos son categorías.

Un concepto constituye una categoría puesto que permite delimitar un conjunto de objetos (físicos o no físicos) que pertenecen a dicha categoría (i.e. constituyen su extensión, son lo denotado por el concepto). Las aplicaciones de un concepto poseen evaluación semántica.

3. Los conceptos poseen composicionalidad.

La idea es que los conceptos pueden combinarse entre sí y que el valor semántico del compuesto es función de los valores semánticos de los constituyentes. Fodor considera que la composicionalidad es la única explicación plausible de propiedades del pensamiento como la sistematicidad y la productividad (cfr Fodor & Pylyshyn, 1988). Su argumento establece un paralelismo entre el caso del pensamiento y el caso del lenguaje: la sistematicidad del lenguaje sólo puede explicarse atribuyéndole una estructura composicional; pero el pensamiento también es sistemático, por tanto, el pensamiento también debe ser composicional, i.e. el pensamiento, como el lenguaje, debe poseer una estructura de constituyentes.

4. Los conceptos (muchos de ellos) pueden ser aprendidos.

Adquirir un concepto, argumenta Fodor, implica quedar trabado (*locked*) nomológicamente a la propiedad que el concepto expresa. De este modo no es necesaria una hipótesis mediadora en la que se despliegue el concepto. El concepto así adquirido puede ser a un tiempo primitivo y no innato. Esta es la premisa que le permite a Fodor concluir que la mayoría de los conceptos son primitivos, aunque no figuren en el repertorio inicial de la estructura conceptual del bebé. Esto es, Fodor rechaza las teorías de primitivos semánticos, las teorías que sostienen que existe un conjunto relativamente poco extenso de elementos semánticos a los que es posible reducir la totalidad de los significados de todos los conceptos.

5. Los conceptos son públicos y compartidos.

La teoría representacional de Fodor hace uso de la distinción tipo/caso y de la noción de realizabilidad múltiple: existen particulares mentales (casos) que instancian el mismo concepto (tipo). Así, dos personas diferentes pueden poseer el mismo concepto instanciado a través de dos casos que pertenecen al mismo tipo. La idea de Fodor es evitar aquí el fantasma del relativismo. Aunque dos personas difieran en buena parte de sus experiencias (físicas,

culturales o de otra clase) con respecto a perros o triángulos, eso no impide que ambos posean el respectivo concepto PERRO o TRIÁNGULO. Fodor se aferra a la idea de que existe alguna noción de identidad lo bastante robusta como para poder ser aplicada a los conceptos. La noción de similaridad no puede suplantar a la de identidad, ya que siempre presupone a esta última.

La cuestión que ahora se nos plantea es cuál(es) de las condiciones puede abandonar el teórico del CNC. Sin embargo, los problemas de éste no se acaban aquí. Desde la óptica de Fodor las condiciones son falibles pero no negociables, en el sentido de que si abandonamos cualquiera de ellas, el precio que debemos pagar es la propia teoría representacional de la mente. De modo que nos encontramos ante el siguiente dilema: o bien el CNC cumple todas las condiciones, con lo cual resuelta indistinguible del CC, o bien hay alguna que no cumple, en cuyo caso no es representacional. El dilema tiene el clásico regusto fodoriano de un juego en el que si se aceptan las reglas no es posible ganar, puesto que en realidad se están ofreciendo idénticos criterios para “conceptual” y “representacional”. No obstante, pone el dedo en la llaga al señalar que el teórico del CNC debe presentar una noción de representación alternativa.

¿Qué condiciones abandona una teoría del CNC? Descartemos en primer lugar la condición 1: abandonarla no marca el contraste entre CC y CNC, sino entre una orientación realista y una de corte instrumentalista. La condición 4 también se mantiene: es muy implausible que el teórico del CNC quiera mantener que todas las representaciones con este tipo de contenido tienen carácter innato. Bermúdez explícitamente mantiene las condiciones 2 y 3: el CNC es evaluable semánticamente dado que permite la representación falsa³ y sería así mismo composicional. La condición 5 se presenta entonces como la candidata a discriminar entre CC y CNC.

¿Es el CNC público y compartido?

En primera instancia, parece que la condición que el teórico del CNC puede abandonar es la referida a la naturaleza pública y compartida. El CNC es contenido tal y como se presenta en la experiencia. El modo en que experimentamos la realidad, puede argumentarse, admite un cierto grado de idiosincrasia. En consecuencia, ante una idéntica porción de realidad dos sujetos no instancian necesariamente el mismo CNC. Este planteamiento

³ Para Peacocke, (1992) el CNC cumple condiciones de corrección. Sin embargo, resulta harto difícil determinar dichas condiciones sin recaer en la noción de evaluabilidad semántica establecida en la condición 2 de Fodor (de un modo semejante a lo que ocurre con la noción de adecuación en McGinn (1989), véase más adelante).

resulta problemático para los fines fundacionales del CNC. Si las representaciones con CNC resultan idiosincráticas, ¿cómo es posible que puedan proporcionar los elementos con los que se construyen los conceptos comunes entre diferentes individuos? En este punto uno puede seguir las conclusiones hasta el final e insistir, como Stalnaker (1997), en que no existe identidad interindividual en el mundo de lo conceptual. Stalnaker realiza un movimiento de cierre similar al de McDowell (1994) pero en la dirección opuesta. Si McDowell concibe la esfera de lo conceptual como algo que abarca el ámbito de lo experiencial⁴, Stalnaker toma como punto de partida lo no conceptual para extenderlo hacia el mundo de los conceptos. La consecuencia es que nuestros conceptos pueden diferir en la medida en que nuestras experiencias lo hagan, y la brecha será tanto mayor cuanto más alejados se encuentren nuestros repertorios experienciales.

Una solución distinta es la de Cussins (1990), para quien la objetividad, i.e. la posibilidad de representar objetos y propiedades que pueden ser sujetos a condiciones veritativo-funcionales, es algo que emerge a partir de las experiencias preobjetivas. La capacidad conceptual depende de la estructura no conceptual del contenido de las capacidades cognitivas (como las capacidades de discriminación y de seguimiento espacio-temporal de las entidades del mundo). Los conceptos atómicos poseen una estructura no conceptual que ha sido modelada en la experiencia. La estructura combinatoria de los conceptos, su sintaxis y semántica, puede captarse por medio de la estructura representacional no sintáctica y no semántica interna a los conceptos. En este esquema, es el CNC el que lleva a cabo la verdadera labor explicativa. Lo conceptual es una construcción a partir del CNC, pero la eficacia psicológica pertenece a este último. A pesar de que en las representaciones de tipo lingüístico es el contenido conceptual el que resulta más aparente, es la estructura subyacente no conceptual la que explica este contenido. Los conceptos, señala Cussins, son en realidad idealizaciones de esta dicha estructura.

Además del problema del papel irrelevante que los conceptos desempeñan en la vida mental, esta posición sigue sin poder dar cuenta de su naturaleza compartida. Nada garantiza que se lleve a cabo la misma idealización

⁴ El propósito de McDowell es reconducir la situación hacia la reconciliación sugerida por el lema kantiano "los pensamientos sin contenido son vacíos; las intuiciones sin conceptos son ciegas" (cfr. *Crítica de la razón pura* A51/B75). El problema es cómo conseguir que pensamientos e intuiciones se encuentren conectados racionalmente. La solución que encuentra consiste en entender el mundo de lo conceptual como algo que se extiende hacia la experiencia. Las experiencias son pasivas pero reflejan capacidades conceptuales. Las impresiones en los sentidos poseen ya contenido conceptual pero eso no impide que sean experiencias *del* mundo.

entre sujetos. Sin embargo, una visión alternativa es posible si atendemos a que las experiencias preobjetivas pueden ser idénticas entre individuos y pueden venir constituidas por interacciones con elementos públicos. En realidad, es sostenible que los aspectos idiosincráticos a la hora de captar la realidad pueden originarse en los distintos modos de conceptualizarla que poseen los sujetos, no en el hecho de que vengan contruidos desde experiencias inherentemente idiosincráticas. El carácter compartido de dichas experiencias puede venir garantizado (dentro de ciertos límites) por ciertos patrones de interacción entre el individuo y el entorno.

La posición de Bermúdez (1998) puede entenderse de esta manera, al apelar a mecanismos universales de interacción con el ambiente y con otros individuos. Consideremos dos de los ejemplos que propone: En su explicación del marco en el que aparece la concepción del sí-mismo en un niño preverbal, Bermúdez emplea con profusión elementos que no se avienen a una lectura subjetivista. Uno de ellos es la noción de *affordance* (Gibson, 1979), que Bermúdez emplea, entre otras cosas, para dar cuenta de la información que un sujeto precisa para construir un sentido del “sí mismo” previo a la adquisición del concepto YO. Una *affordance* es una regularidad de alto nivel que el sujeto, de acuerdo a Gibson, es capaz de percibir directamente. Una *affordance* se caracteriza por ser intraespecífica por lo cual cabe suponer la existencia de mecanismos comunes entre los miembros de una especie para captarla. Es más, la percepción de *affordances* tienen todos los visos de ser innata. En consecuencia, cabe suponer que el contenido representacional basado en tales propiedades posee una naturaleza pública y compartida determinada por la regularidad de determinadas propiedades del entorno y por la regularidad de los mecanismos cognitivos que captan dichas propiedades.

Idénticas observaciones se pueden realizar respecto al fenómeno de la *atención visual conjunta*, i.e. las interacciones en las que el padre dirige la mirada del hijo (o viceversa) hacia objetivos específicos. Bermúdez señala que la mejor explicación de este fenómeno requiere atribuir al niño la capacidad de reconocerse en cuanto sujeto de percepciones, y al mismo tiempo reconocer que el padre es igualmente capaz de percibir. Sin embargo, Bermúdez omite un tercer factor necesario para hacer inteligible la interacción: la capacidad del niño para reconocer que el otro “ve lo que veo yo”. Esto es, debemos incluir el supuesto tácito en las personas que interaccionan de que no existen diferencias sustanciales en sus respectivas experiencias del objeto en cuestión. Al atraer la atención del otro hacia un objetivo, el sujeto consigue hacer pública su propia experiencia, y permite que el otro la comparta.

Recapitulando, las teorías del CNC que inciden en su carácter no público tienen problemas para llevar a cabo un proyecto de construcción de conceptos que no desemboque en el relativismo; y teorías como las de Bermúdez emplean nociones que hacen dudar del carácter realmente no público del CNC que postulan. Por otra parte, si el CNC cumple un criterio de corrección, entonces hay un sentido en el que dos casos de CNC pueden poseer las mismas condiciones de corrección e instanciar así el mismo tipo, con lo cual podemos hablar del mismo CNC en distintos sujetos. ¿Existe alguna salida que evite que el CNC resulte indistinguible de los conceptos de Fodor? A continuación voy a explorar brevemente una alternativa más drástica, como es rechazar la condición 2.

La interdependencia de CC y CNC

Abandonar la condición 2 de Fodor entraña que las entidades con CNC no poseen realmente evaluación semántica, no podemos decir de ellas que se refieren, denotan, etc a X. Esto parece tener consecuencias catastróficas para el teórico del CNC, puesto que se trata de la condición más estrechamente ligada a la representacionalidad de aquello que postula. Una entidad que no cumpla esta condición no puede constituir una representación falsa (*misrepresentation*) lo cual para muchos (incluido Bermúdez, 1998, p. 84) entraña que no puede ser una representación en absoluto. No sólo eso, sino que la composicionalidad parece quedar amenazada: el contenido de los compuestos no puede ser función del contenido de las partes, si las partes no poseen contenido del que partir.

El teórico del CNC puede intentar desarrollar una teoría de la representacionalidad que prescinda de la idea de ‘representación falsa’. Una alternativa de este tipo es la que defiende McGinn (1989), a través de la idea de modelos mentales⁵. Un modelo mental es una estructura mental que simula un determinado aspecto de la realidad. McGinn sostiene que, a diferencia de las representaciones de base lingüística, un modelo mental no posee propiedades semánticas: no es verdadero o falso acerca de la realidad que se refleja en él. No obstante, el modelo recoge la realidad de un modo adecuado o inadecuado, con mayor o menor verosimilitud. Las representaciones basadas en un modelo adecuado vendrán etiquetadas como verdaderas y las basadas en un modelo inadecuado como falsas. Esta visión presenta obvios problemas, quizás el más evidente el de encontrar un noción de “adecuación” que no nos remita de nuevo a la evaluación semántica. McGinn parece tener en mente un criterio de utilidad

⁵ Un modelo mental no tiene por qué ser una representación con CNC, aunque esto no queda claro en la teoría de McGinn.

adaptativa, de modo que los modelos más adecuados sean aquellos que otorgan al agente mejores opciones de enfrentarse a los retos del ambiente. En muchas ocasiones tales modelos serán los que capturen más fielmente los aspectos relevantes del entorno. Pero en muchas otras las representaciones falsas de la realidad pueden resultar adecuadas desde el punto de vista adaptativo, por ejemplo, aquellas que se dan en las falsas alarmas (i.e. el animal que huye al detectar erróneamente un predador), en las que es normalmente más adaptativo errar del lado de la prudencia. Por otra parte, no es necesaria la representación para cumplir con una condición de adecuación teleológica. Los modelos mentales podrían cumplir la función correspondiente y ser, sin embargo, entidades puramente informacionales.

Si resulta problemático dar con una teoría de la representación que prescinda de la condición de evaluación semántica, ¿queda alguna alternativa que permita salvar la representacionalidad del CNC sin transformarlo en CC? En mi opinión la hay, pero requiere que abandonemos el PA. La idea básica es considerar a CC y CNC como interdependientes: para que una entidad adquiera CNC se precisa la presencia de un concepto; para que una entidad adquiera CC se requiere experiencia no conceptual. Si, de un modo aproximado, caracterizamos el CNC como el aspecto en el que se nos ofrece la experiencia, lo que el concepto permite es delimitar dicha experiencia, hacerla experiencia de *algo*. El CNC, por su parte, ancla el concepto en el mundo, hace al concepto representación *de algo*.

Para exponerlo de otro modo: Un estado con contenido es un estado representacional, y un estado representacional es informacional, pero no todo estado informacional es un estado representacional y, en consecuencia, con contenido. Para ello un estado informacional debe poseer una métrica. En ausencia de una métrica, el (putativo) CNC no se constituye como representación, esto es, como objeto intencional. Si el CNC proporciona su propia métrica, entonces resulta indistinguible del CC. Si existe una métrica para el CNC ha de ser extrínseca a él. La métrica la proporciona el concepto que se instancia (o más propiamente: la historia de conceptos que se han instanciado) cuando se instancia un determinado CNC. El CNC no categoriza, si no es a través de un concepto. Pero el concepto sólo categoriza la realidad en cuanto presentada por un CNC determinado.

McGinn se acerca a esta idea cuando se refiere a las relaciones entre proposiciones y modelos mentales. Una proposición puede indexar un modelo mental, en el sentido de que le asigna un elemento de un determinado espacio lógico. Lo que sugiero es que lo conceptual cumple una función de índice similar respecto a lo no conceptual, pero una función que no limita el concepto

a ser una glosa del CNC, sino que le permite a éste constituirse como representación, le otorga carta de intencionalidad al referirlo a una porción de realidad determinada y no otra. El CNC, por su parte, permite “llenar” el concepto, conectarlo con una estructura que preserva análogamente la estructura de la realidad. De este modo nos deshacemos del fantasma de un concepto que, por emplear una expresión de McDowell, “gira sin freno en el vacío”. Señalábamos mas arriba que para Peacocke CC y CNC deben elucidarse simultáneamente. Lo que apunta la interdependencia es que se *constituyen* simultáneamente.

Es necesario, así mismo, diferenciar la noción de interdependencia de la posición de McDowell. Del mismo modo que al decir que existe una relación de dependencia de un objeto mental respecto al entorno no habríamos de entender que el entorno forme parte constitutiva de dicho objeto, tampoco al decir que la intencionalidad de un determinado objeto depende de un determinado concepto estamos dando a entender que el objeto se constituya conceptualmente. Lo único que decimos es que dicho concepto o dicho entorno forman parte de las condiciones para que el objeto en cuestión se constituya. Pero estamos hablando de *interdependencia*, por tanto el concepto no se instancia si no es a través de un determinado CNC. Lo cual no significa que el concepto sea reducible al CNC, esto es, que el CNC constituya la estructura interna del concepto⁶. Contrastemos tres posibles historias para describir el caso de un recién nacido que mira un rostro:

a) Prioridad de lo no conceptual: el niño posee un cierto tipo de experiencia, que al acumularse con (y en relación a) un cierto número de experiencias similares en contextos diversos, conducen a la adquisición del concepto ROSTRO.

b) Prioridad de lo conceptual: el niño posee un concepto (ROSTRO) en torno al cual se organiza y estructura la experiencia. La experiencia es conceptual.

c) Interdependencia: el niño instancia un concepto (ROSTRO?) ante ciertas experiencias. A partir de ese momento, experiencias similares son experiencias de rostros. Pero las experiencias no son conceptuales: las experiencias no son ROSTRO.

Una consecuencia de la interdependencia es lo que podemos llamar *innatismo conceptual mínimo*. La idea de innatismo mínimo no es nueva: Clark (1993) defiende que los mecanismos mentales con los que cuenta el recién nacido poseen sesgos innatos que le permiten conducir el desarrollo de las estructuras funcionales en las direcciones apropiadas. Lo que se añaden

⁶ Es posible, como señala Fodor (1998), que un concepto no posea en realidad estructura.

‘conceptual’ es que dichos sesgos los proporciona una estructura de conceptos (o, más propiamente expresado: de *conceptualizadores*) que es mínima en un doble sentido: en que probablemente está orientada a un pequeño conjunto de aspectos de la realidad (aquellos más importantes para la supervivencia), y en que los conceptos no se constituyen como entidades intencionales propiamente dichas hasta que no son llenados por el CNC, a medida que el niño acumula experiencias del mundo. Si pudiésemos “congelar” la mente del recién nacido en un estado inicial S_0 (previo a toda experiencia) no encontraríamos realmente contenido, sólo organizadores apuntando en las direcciones correctas para que ese contenido llegara alguna vez a instanciarse.

Las condiciones de formación del concepto (o percepto, asumiendo que éste es el correlato del CNC, lo mismo que el concepto lo es del CC) no forman parte de sus condiciones de subsistencia. Dicho de otro modo, asumiendo que existe una estructura física (un vehículo) que instancia cada contenido en cuestión, una vez que dicha estructura ha adquirido su representacionalidad, no la pierde. Que la constitución de un determinado concepto (v.g. MAMÁ) requiera de un determinado cúmulo de experiencias (experiencias de mamá), no significa que el concepto no pueda desplegarse de modo independiente a dichas experiencias (e inversamente para los correlatos del CNC). (Como señalaré un poco más adelante, aunque CC y CNC van habitualmente de la mano es posible que se encuentren disociados en circunstancias especiales). Un modo de ver desde otra perspectiva la idea de interdependencia es considerar algunos posibles *vehículos* de las representaciones, i.e. las entidades computacionales que podrían sostenerlas.

CC, CNC y vehículos representacionales

Aunque no es preciso ligar la relación entre CC y CNC a un modelo computacional determinado, no es ajeno a las teorías del CNC relacionarlas con las posibilidades que ofrecen los sistemas conexionistas (cfr Clark, 1993; Chrisley, 1993; Peacocke, 1994). Las sugerencias giran a menudo en torno de cómo caracterizar ciertos estados de las redes conceptualmente a partir de estados que poseen CNC. Esto es, se esboza una explicación en que los estados representacionales de la red son inicialmente no conceptuales para posteriormente integrarse en estados (más estables) que constituyen el concepto. Sin embargo, como señala Ramsey (1997) resulta problemático motivar la representacionalidad de tales estados, en vez de una descripción puramente informacional, sin hacer omnipresente y vacua la noción misma de representación.

Un ensayo de solución requiere: negar que los estados iniciales de la red son estados con contenido; mostrar cómo cualquier contenido que se atribuya a la red debe atribuirse a ésta en su totalidad; mostrar cómo el contenido es función de la historia de entrenamiento de la red y la presencia de un organizador que determina cuál es la medida de error respecto a la que juzgar cuándo la red se acerca a determinado objetivo. En la mayoría de los casos este organizador lo proporciona el investigador que experimenta con la red. La sugerencia es que en el caso natural el organizador es un elemento interno presente desde el estado inicial del sistema. O dicho de otra manera: un concepto (mínimo, en el sentido introducido más arriba) innato. El concepto, sin embargo, no se constituye como tal (no tiene contenido, no es concepto de X) sino una vez que el sistema entra en una dinámica de contacto con la experiencia. Al igual que en el recién nacido, si fuera posible “congelar” el sistema en un estado inicial S_0 , estaría completamente vacío de contenido: los organizadores no son capaces de proporcionarlo por sí mismos⁷.

Disociando CC y CNC

La relación tan estrecha que la interdependencia plantea entre lo conceptual y lo no conceptual se presta fácilmente a una objeción. Si ambos se requieren mutuamente para constituirse como entidades con contenido, ¿no habremos acabado cayendo en la identificación entre ambas esferas que queríamos evitar? ¿No será que, a fin de cuentas, *eso* con CNC y *eso* con CC que se constituyen simultáneamente son una única y sola entidad? Consideremos el ejemplo de la red y el “organizador”: ¿no puede interpretarse que el conjunto “red + organizador” constituye en realidad el concepto? Por otra parte, ¿tiene alguna trascendencia la distinción para la teorización acerca de los fenómenos mentales?

Ya se ha señalado que el que X dependa de Y para constituirse como objeto intencional, no entraña que Y sea constitutivo de X. Tampoco que Y deba estar presente las sucesivas ocasiones en que X se instancia. Una vez constituido, podemos decir que X goza de cierta independencia de Y: el

⁷ Las redes distribuidas también nos permiten atisbar una solución al abandono de la composicionalidad que se sigue del abandono de la condición 2. Dejando a un lado las lecturas de las redes en términos de microrrasgos (que son susceptibles de interpretarse como “microconceptos”), es posible ver cómo el contenido de la red (global) depende en forma sistemática de sus elementos (unidades y conexiones), a pesar de que estos carecen de contenido. Ello permitiría contemplar transiciones de contenido a contenido vía forma, pero sin que cada uno de los elementos formales posea contenido a su vez (cfr Horgan & Tienson, 1996). En cualquier caso, es dudoso que el conexionismo pueda dar respuesta al reto de Fodor & Pylyshyn (1988) y ofrecer una explicación de propiedades como la sistematicidad que no descansen en una misma ontología (i.e. una ontología conceptual).

concepto no necesita del percepto para referirse a lo que se refiere y viceversa. Por otra parte, existen fenómenos mentales que tal vez puedan contemplarse en términos de una disociación entre CC y CNC. Bermúdez (1998, p. 81) sugiere ciertos tipos de agnosia visual (las asociativas) son caracterizables en términos de daño selectivo en el nivel conceptual, con el CNC relativamente intacto, mientras otras agnosias (las aperceptivas) implican un daño en el nivel no conceptual, con el CC relativamente intacto. Si esta interpretación es correcta, la consecuencia es que el CNC se emplea en la construcción de conceptos pero no en las sucesivas instanciaciones de los conceptos (dado que las agnosias asociativas los preservan a pesar del daño en el CNC).

Un segundo ejemplo, sugiero, lo constituyen ciertos estados esquizofrénicos alucinatorios y delusorios. Como señala Sass (1992, p. 109), a diferencia de otros desórdenes psicóticos las alucinaciones y delusiones esquizofrénicas resultan virtualmente incomprensibles debido a que implican alteraciones en las estructuras más fundamentales (tiempo, espacio, identidad humana). La dificultad para caracterizar la experiencia de estos sujetos le lleva así a echar mano de la distinción heideggeriana entre lo óntico y lo ontológico: el problema radicaría no tanto en que el sujeto construye entidades incorrectas, sino en que el estatuto ontológico de dichas entidades es puesto en entredicho. Mi sugerencia (especulativa, sin lugar a dudas) es que quizás sea posible dar un tratamiento a tales fenómenos en términos de CC y CNC, lo cual permitiría una integración más suave del estudio de dichos fenómenos con las herramientas descriptivas y explicativas de la ciencia cognitiva. A diferencia de las agnosias no parece que en estos casos uno de los niveles presente problemas mientras el otro permanece intacto, sino en algún tipo de “desconexión” entre ambos: el esquizofrénico podría tener la mente literalmente dividida en un mundo de representaciones experienciales que no engarza con el mundo de representaciones conceptuales. No es sorprendente, entonces, que la ingesta de sustancias psicoactivas, como el LSD, cuyos síntomas son a menudo descritos como similares a los de un estado psicótico, sea descrita por el individuo en cuestión como un intento de captar “la experiencia en estado puro”. La dificultad de describir estas experiencias sugiere una vez más el estrecho vínculo que es habitual entre CC y CNC.

Conclusiones

La noción de contenido no conceptual es una idea intrigante que precisa un mayor desarrollo para hacerla inteligible. En esta comunicación he considerado algunas de las dificultades para distinguirla del contenido conceptual, especialmente para evitar que sea fagocitada por este último. Un

problema capital es dar cuenta de la representacionalidad: si lo no conceptual se presenta como representacional y autónomo, o bien se abre una brecha que impide el proyecto de fundamentar en él la esfera conceptual, o bien se construye de tal manera que resulta indistinguible de esta última. El principio de autonomía, por tanto, es difícilmente sostenible. Ambos contenidos mantienen una relación de interdependencia y estudiar las bases de esta relación, así como su sustrato computacional, podría proporcionar alguna pista de cara a comprender ciertos intrincados fenómenos patológicos⁸.

BIBLIOGRAFÍA

- Bermúdez, J. L. (1994) "Peacocke's Argument against the Autonomy of Nonconceptual Content", *Mind and Language*, 9: 403-418.
- Bermúdez, J. L. (1998) *The Paradox of Self-Consciousness* Cambridge, MA: MIT Press.
- Chrisley, R. L. (1993) "Connectionism, Cognitive Maps, and the Development of Objectivity", *Artificial Intelligence Review*, 7: 329-354.
- Clark, A. (1993) *Associative Engines* Cambridge, MA: MIT Press.
- Cussins, A. (1990) "The Connectionist Construction of Concepts". In M. Boden (ed.) *The Philosophy of Artificial Intelligence* Oxford: Oxford University Press.
- Davies, M. (1998) "Language, Thought and the Language of Thought (Aunty's Own Argument Revisited)". In P. Carruthers & J. Boucher (eds.) *Language and Thought: Interdisciplinary Themes* Cambridge: Cambridge University Press.
- Dretske, F. (1995) *Naturalizing the Mind* Cambridge, MA: MIT Press.
- Evans, G. (1982) *The Varieties of Reference* Oxford: Oxford University Press.
- Fodor, J. (1998) *Concepts* New York: Oxford University Press.
- Fodor, J. & Pylyshyn, Z. (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, 28: 3-71.
- Gibson, J. J. (1979) *The Ecological Approach to Visual Perception* Boston: Houghton Mifflin.
- McDowell, J. (1994) *Mind and World* Cambridge: Harvard University Press.
- McGinn, C. (1989) *Mental Content* Oxford: Blackwell.
- Peacocke, C. (1992) *A Study of Concepts* Cambridge, MA: MIT Press.
- Peacocke, C. (1994) "Non-conceptual Content: Kinds, Rationales and Relations", *Mind and Language*, 9: 419-430.
- Ramsey, W. (1997) "Do Connectionist Representations Earn Their Explanatory Keep?", *Mind and Language*, 12: 34-66.
- Sass, L. A. (1992) "Heidegger, Schizophrenia, and the Ontological Difference", *Philosophical Psychology*, 5: 109-132.
- Stalnaker, R. (1997) "What Might Nonconceptual Content Be?", paper presented at the SOFIA Conference, Barcelona, June 1997.

Fernando Martínez Manrique
Rutgers University
e-mail: fmartine@rci.rutgers.edu

⁸ Esta comunicación ha sido realizada en el marco de una beca postdoctoral BFI98.91 del Departamento de Educación del Gobierno Vasco.

SHOEMAKER Y EL PROBLEMA DE LOS *QUALIA**

MARIANO RODRÍGUEZ GONZÁLEZ
Universidad Complutense de Madrid

ABSTRACT: Shoemaker's compatibilism accepts the realism of *qualia*, binding it to the vindication of the functionalist theory, in a subtle essay on reconciling manifest and scientific images of the world. In this sense, the most important contribution of his will be the coining of the *quale* notion from the experience of qualitative similitude, a strategy by virtue of which our philosopher gets into a relatively objective dimension in order to ground the *qualia*. Undertaking the possibility of inverted spectrum but rejecting the absent *qualia* hypothesis, Shoemaker makes us to see how the compatibilism retreats from the problems of consciousness and subjectivity.

KEYWORDS: "*Qualia* are not easy to domesticate"

I

La noción de funcionalismo que tiene presente nuestro autor se puede decir que sigue la estela del ya clásico trabajo de Block, utilizando igual que él la formulación que debemos a D. Lewis y que recurre a la sentencia-Ramsey de una teoría psicológica T. De lo que se trataría es de poner a punto definiciones funcionales de los estados mentales, sin duda, pero lo importante ahora es no descuidar la distancia que separa a las definiciones *fuertes* de las débiles. Y es que el conductismo lógico sucumbió al círculo de lo mental, mientras que el funcionalismo, para Shoemaker su sucesor más respetable, lo evitará sólo si da expresión a la propiedad mental mediante un predicado funcional que no contenga terminología mentalista. Lo que se requiere entonces son definiciones funcionales fuertes de los estados mentales (1981/1984, 264), y para lograrlas emplearemos la sentencia-Ramsey de una teoría psicológica (Block 1980, nota 1) que para Lewis se identifica con la *folk psychology*. Se nos puede ocurrir que con ello estamos en realidad ante la enésima estrategia de eliminación de lo mental: definimos el dolor aludiendo,

* Este trabajo se ha realizado bajo el estímulo del Proyecto de Investigación MEC-PB96-0580/96 n°6, financiado por la D.G.I.C.Y.T.

entre otras cosas, al deseo de librarse de él, pero a uno le duele algo, en definitiva, si y sólo si Existe algún $F_1 \dots F_n$ [$T (F_1 \dots F_n) \& x$ tiene F_i]. Y sin duda no es descabellado afirmar que el principio del funcionalismo de Lewis, el llamado *analítico*¹, según el cual el estado mental M se identifica por definición con el estado que causan típicamente $C_1, C_2, C_3 \dots C_n$, y con el estado que típicamente da como resultado las conductas $B_1, B_2, B_3 \dots B_n$, es un principio conductista (al menos) en su origen. Pero Shoemaker contestará a esta objeción que el *definiens* de una definición correcta de estados mentales, por mucho que no contenga términos referidos a estados mentales *particulares*, cuantificará sobre lo que de hecho son estados mentales (311). Y ya comprobaremos el eco que esta respuesta tendrá en su compatibilismo.

Tesis fundamental de todos los funcionalistas es la de la múltiple realizabilidad física del mismo estado mental, o, como lo pone nuestro autor, que tener una realización total dada de una propiedad funcional es condición suficiente pero no necesaria para tener la propiedad. Ésta puede instanciarse en virtud de la instanciación de alguna realización total muy diferente (265). Está clara entonces la vocación materialista del funcionalismo a pesar de los titubeos de los inicios: podemos identificar una propiedad funcional con la disyunción, tal vez infinita, de todas sus realizaciones totales, y considerar, como desde luego hace Shoemaker, que todas estas realizaciones son *efectivamente* físicas. *A priori* ni siquiera el materialista podría descartar realizaciones posibles no físicas, pero le llega con sostener que en el mundo actual se trata siempre sólo de *physical disjuncts*. Por eso hay que afirmar que el compatibilismo de Shoemaker constituye una versión más, particularmente sofisticada, del proyecto de conciliar las imágenes manifiesta y científica del mundo. Habían puesto el dedo en la llaga Block y Fodor al señalar que la nueva teoría de la mente es incapaz de acomodar el carácter cualitativo o fenomenológico de los estados psíquicos. Porque con ello nos hacían reparar de nuevo en la herida abierta en el pensamiento occidental desde la revolución científica moderna. Y es que nada sería un caso del tipo dolor si no lo sintiéramos como dolor: ¿en qué quedaba entonces el tan alabado punto fuerte de los funcionalistas, esa identificación funcional que aspiraba a hacer los estados mentales científicamente manejables?. Hasta cabía establecer la posibilidad de dos de esos estados funcionalmente equivalentes pero uno con carácter cualitativo y el otro sin él. A la altura de *Reason, Truth and History* el mismísimo Putnam reconocería que experimentar una tonalidad particular de

¹ Del estudio sobre las variedades del funcionalismo se puede extraer la conclusión de que nuestro autor, no obstante su declaración explícita, acaba encontrando razones para preferir a los demás el funcionalismo analítico, sobre todo si lo combinamos con lo que denomina "funcionalismo mínimo" o TCP. Y es que el Psicofuncionalismo desembocaría en el "parroquialismo". Es justamente *la comprensión que tenemos de nuestro vocabulario mental* la que excluiría la idea de que los extraterrestres no tienen dolor o deseos al no compartir nuestra psicología profunda.

azul le obligaba a uno a retroceder a la vieja teoría de la identidad (1981, 87-88)

II

Galileo y Locke negaron que las cualidades secundarias inhieran en las cosas externas, como nos parece percibir. Lo que sucede es que nuestras sensaciones tienen *propiedades subjetivas* inmediatamente accesibles a la conciencia, que no “se parecen” a las propiedades objetivas exteriores. Tal y como los experimentamos, los colores existirían “sólo en la mente”. Pues bien, para hablar de estas propiedades intrínsecas el filosofar reciente emplea el término de *qualia*, sin poder recurrir ya al cómodo expediente de relegarlas al mundo de lo mental, en vías de ser colonizado. Y nuestro autor va a insistir en que este aspecto fenoménico de las sensaciones y de los estados perceptivos es precisamente lo que determina *cómo se siente uno al tenerlas* (1991, 333). Es por tanto el carácter subjetivo de los estados conscientes lo que daría sustancia a la cuestión de los *qualia*, de ahí que la estrategia de echar mano de la teoría de la identidad para resolverla no nos parezca muy eficaz.

Además hay que reseñar que los *qualia* se supone que son rasgos no intencionales de las experiencias: una cosa es el contenido representacional y otra distinta el contenido cualitativo que subyace a aquél, que no estaría explícitamente representado en nuestro experimentar. Shoemaker se dirige contra el modelo de introspección que está tomado de la percepción de objetos, según el cual conocemos *hechos* a partir de la percepción de *objetos* y de sus relaciones mutuas (1994, 21). Esto implica negar, contra Harman, que las únicas propiedades accesibles a la introspección sean las intencionales, y además conlleva el rechazo de la teoría de los *sense data*. El color de la hierba está donde lo vemos, acaba declarando nuestro autor, lo que ocurre es que en cuanto fenómeno se halla constituido por la estructura (y las limitaciones, a la kantiana) de nuestro equipo visual, de nuestro *espacio cualitativo*, para decirlo con las palabras de Quine.

Pero la noción de *quale* es muy inestable porque resulta completamente “transida de teoría”. Por eso Shoemaker recurrirá para acabar de perfilarla a la experiencia de la *semejanza cualitativa*, algo que ya apuntaron Quine y Putnam. Será definida ésta en un primer momento desde aquélla, es decir, los *qualia* no serían sino los rasgos fenoménicos de las experiencias sensoriales en virtud de los cuales unas se parecerán y diferirán cualitativamente de otras (1981/1984, 336). Es un hecho bruto que los estados sensoriales se parecen unos a otros y se diferencian unos de otros. Estas relaciones las mantienen en virtud de sus rasgos fenoménicos, está claro. Ahora bien, la experiencia de la semejanza cualitativa sí que podemos decir que nos ancla en un mundo objetivo, puesto que tiene que darse una “conexión legaliforme” entre el hecho de que una persona tenga experiencias

fenomenológicamente similares y su creencia de que existen *similitudes objetivas* en el mundo (1975/1984, 180-1). Así que es evidente la ventaja de partir de la experiencia de la semejanza fenoménica, de forma que en un segundo paso nuestro autor procederá a acuñar el mismo concepto de *quale* desde ella². Podemos definir la clase de los estados cualitativos sencillamente porque los estados mentales mantienen entre sí semejanzas y diferencias fenoménicas que no son las semejanzas y diferencias intencionales (y espero que ahora esta última distinción se nos haga un poco más comprensible).

Es así la similitud cualitativa la que nos llevaría a suscribir, frente a Dennett o al ya citado Harman, *el realismo de los qualia* que defiende nuestro autor, “la necesidad de tomarse los *qualia* en serio”. Ocurre que las propiedades microestructurales asociadas con cada color son muy diversas. La rojez, por ejemplo, la podemos identificar con una propiedad física, pero disyuntiva en alto grado. Y lo que confiere unidad a los disyuntos no tiene que ver con una supuesta naturaleza física intrínseca que los convierta en “clase natural”, sino que consiste más bien en su relación con criaturas dotadas de sistemas perceptivos como el nuestro. Propiedades microestructurales diversas serán “realizaciones” de un mismo aspecto cromático si son semejantes sus efectos en esos sistemas perceptivos. Lo que también vale para la similitud y la diferencia cromáticas, evidentemente, de manera que aquellos que posean espacios cualitativos alternativos no estarán percibiendo erróneamente una semejanza o diferencia objetiva entre colores, sino que habrá que considerar esas relaciones relativas al observador. Nuestras experiencias perceptivas, si queremos decirlo así, proyectan inevitablemente en el mundo la estructura de nuestro espacio cualitativo, que es cuestión de tener la capacidad, en buena medida innata, de discriminar entre diferentes estímulos sensoriales. Pero reconocer esto no nos tiene que comprometer con ninguna forma tradicional de proyectivismo, dado que Shoemaker ha encontrado la forma de anclar objetivamente, como dijimos, los aspectos fenoménicos de las experiencias a través de la semejanza cualitativa entre las mismas.

III

¿Por qué habría de ser descartada por inviable una definición funcional del dolor?. No sólo podríamos mencionar las tendencias a influir en la conducta abierta y a generar en la persona que lo padece la creencia de que algo marcha mal, sino además la propensión a dar lugar a la conciencia

² “Si los estados mentales pueden ser similares o diferentes en *carácter cualitativo*, deberíamos poder hablar de una clase de estados, digamos los *estados cualitativos*, cuyas *condiciones de identidad de tipo* podrían especificarse en términos de la noción de semejanza cualitativa (o *fenomenológica*). Por cada carácter cualitativo determinado que un estado pueda tener, habrá (es decir, podremos definir) un determinado estado cualitativo que una persona tiene sólo en el caso de que tenga un estado que posea precisamente ese carácter cualitativo” (1975/1984, 187)

introspectiva de sí mismo como teniendo un determinado carácter cualitativo, o sea, a producir lo que Shoemaker llama una *creencia cualitativa*: sería sobre todo así como el dolor se relaciona con otros estados psicológicos (1975/1984, 189). Y, por otra parte, ¿con qué fundamento íbamos a negar que la referencia a *qualia* forma parte de una caracterización funcional de la visión?. Que S vea que algo es azul implica, entre otras cosas, que S tiene un repertorio de estados cualitativos que incluye estados asociados con los colores de los objetos, que la estimulación visual bajo ciertas condiciones produce el *quale* asociado, que los grados de semejanza fenoménica entre los *qualia* corresponden a los grados de semejanza entre los colores asociados, que S, por fin, se halla en el estado cualitativo asociado con el color azul (193).

La clave del compatibilismo shoemakeriano está en el descubrimiento de la posibilidad de caracterizar funcionalmente las nociones de similitud y diferencia cualitativas, pues no en vano, como venimos de ver, se ha acuñado a partir de ellas el mismo concepto de *quale*, habiéndosele dotado así además de una relativa objetividad³. Como las condiciones de identidad-tipo de los estados cualitativos se han de especificar en términos de semejanza cualitativa, y nuestro autor parece aproximarse en este punto decisivo a la doctrina russelliana de que los *qualia* son los universales *par excellence*, el hecho de que haya estados mentales con rasgos fenoménicos no tiene por qué plantearle entonces problema alguno al funcionalista. Tener carácter cualitativo no supone para un estado psicológico sino el ser cualitativamente similar a otro estado o estados, o diferente de otro estado o estados, y ya contamos con una explicación funcional mínima de esas semejanzas y diferencias⁴.

Podemos entonces justificar en este mismo momento la importante restricción que el compatibilismo de nuestro autor se autoimpone: puesto que la semejanza cualitativa lo que determina son las condiciones de identidad-tipo, no será posible en absoluto definir funcionalmente *qualia particulares*, ni tampoco legítimo hacer referencia a los mismos en la caracterización funcional de estados psíquicos fenomenológicamente cargados. Esta es la

³ “Mi sugerencia es que aquello que hace de una relación entre experiencias la relación de similitud cualitativa es precisamente el hecho de desempeñar un determinado papel *funcional* en la conciencia perceptiva de similitudes objetivas, esto es, el hecho de que tiende a producir creencias perceptivas al efecto de que tales similitudes tienen lugar” (1975/1984, 199-200).

⁴ Block (1980) se queja de que nuestro autor es muy poco explícito a la hora de mostrarnos cómo una definición funcional de la similitud cualitativa nos ayudaría a definir términos mentales como el de dolor. Lo que sin duda está justificado, pero ni siquiera se aproxima a constituir una auténtica demostración de que el compatibilismo shoemakeriano sea insostenible. Tampoco da en el blanco la reflexión crítica de White (1985), que ciertamente *begs the question* cuando se dirige a convencernos de que las semejanzas y diferencias fenoménicas de que aquí se trata no son genuinas relaciones experimentadas sino tan sólo muertas réplicas funcionalistas sin valor alguno.

razón de fondo, no es sólo que así lo exija también la decisión de aceptar la posibilidad de los *qualia* invertidos. A lo que sí contribuye desde luego la reflexión sobre el espectro invertido es a conformar el compatibilismo moderado según el cual en las definiciones de estados mentales podemos cuantificar sobre *qualia*. Defendiendo esta posición tiene nuestro autor la esperanza de avanzar hacia la reconciliación de dos intuiciones enfrentadas pero muy arraigadas en nosotros: la del fisicalismo y la de la contingencia de las relaciones entre lo mental y lo físico (204). Porque, por mucho que los *qualia* particulares sean “inefables”, lo cierto es que tienen que tener *realizaciones físicas* si damos por sentado que sus condiciones de identidad son funcionalmente definibles. Así que un *quale* podrá ser identificado con la disyunción de sus posibles realizaciones, desde el momento en que cualquier propiedad funcional puede ser físicamente realizada⁵.

En suma, Shoemaker se ha significado por compaginar el realismo de los *qualia* con la decidida adscripción a una concepción funcionalista de la mente, enfrentándose de este modo tanto a los incompatibilistas antifuncionalistas como a los funcionalistas (1991/1996, 121): “Desde un punto de vista materialista, un *quale* será una propiedad de orden superior que una experiencia instancia en virtud de tener una u otra de un conjunto de propiedades físicas pertenecientes a una determinada clase de equivalencia de propiedades físicas” (1996, 152-153). Pero con este resultado parece haberse descuidado esa índole subjetiva de los estados mentales que como vimos son responsables de proporcionar las propiedades cualitativas. Este descuido parece innegable si seguimos a nuestro autor cuando nos dice que en principio podríamos descubrir si las experiencias de los mismos colores en diferentes personas son cualitativamente idénticas o diferentes explorando qué realizaciones físicas tales personas instancian (1996, 145). Aparte de que esto parece dar la razón a los que le acusan de haberse desentendido de la múltiple realizabilidad, nuestro autor no tendrá otra salida que reconocer una vez más que sigue dándose un aspecto de nuestra experiencia, el del *what it is like*, que elude las caracterizaciones funcionales. Que la existencia de este aspecto sea en el fondo compatible con el funcionalismo y el fisicalismo (sea *esperable* a partir de ellos, como subraya Shoemaker) no significa entonces, pensamos nosotros, que se haya resuelto el problema de la disparidad entre imagen manifiesta e imagen científica del mundo, ni el *argument from knowledge*, ni el *explanatory gap*, por mucho que los denodados esfuerzos del pensador americano vayan encaminados a esa meta a través del señalamiento de las condiciones de posibilidad de una posibilidad meramente “aparente”.

⁵ Por esta razón no compartimos la crítica de Doore (1981): la imposibilidad de definir *qualia* particulares no tiene por qué resultar fatal para la teoría funcionalista de mente si la explicación que nos da Shoemaker de la misma se mantiene válida.

IV

Tenemos inversión intersubjetiva del espectro si, por cada matiz cromático, el estado cualitativo asociado en una persona con la visión de ese matiz se halla asociado en otra persona con la visión del inverso de ese matiz. Es muy importante darse cuenta de lo que semejante inversión requiere, a saber, que los espacios cualitativos cromáticos de las dos personas implicadas tengan idéntica estructura, que bajo las mismas condiciones de iluminación ambas formulen los mismos juicios de similitud cromática relativa acerca de los mismos objetos: “A es más parecido a B que a C”. En efecto, no bastaría con que las cosas azules te parezcan a ti como las amarillas me parecen a mí, por ejemplo (1981/1984, 331). En caso de que todo esto fuera posible, no habría manera de decir si la experiencia que tienen de los colores dos personas es la misma o si sus respectivos espectros están invertidos entre sí, porque las dos hipótesis, sin duda, darían lugar a las mismas predicciones comportamentales. En segundo lugar, lo que tendremos será inversión intrasubjetiva si se da un cambio en el modo en que los diversos matices de color se le aparecen a alguien, cada uno de ellos pasando a tener el aspecto que antes de la inversión presentaba su inverso. Es de destacar que este caso intrasubjetivo sí que sería desde luego detectable: el cambio se revelaría a la introspección del sujeto afectado, quien por tanto podría informar del mismo a otras personas.

Block y Fodor utilizaron la posibilidad de los *qualia* invertidos, en el sentido intersubjetivo, para minar la viabilidad del funcionalismo como teoría general de la mente. Pues bien, tras haberlo tenido siempre en cuenta como horizonte de su compatibilismo, admitiendo evidentemente su posibilidad por lo menos lógica, Shoemaker se dispondrá a mostrarnos por qué sólo *parece* que el espectro invertido es incompatible con el programa de la definición funcional de los estados mentales. Y no se trata tanto de proceder a efectuar una aplicación concreta de una doctrina previamente elaborada cuanto de acabar de explicitarla atendiendo a uno de los principales problemas que le dio origen.

Lo primero sería neutralizar la comprensible hostilidad de muchos filósofos, intentando desmentir que la admisión de la posibilidad de la inversión intersubjetiva sea el primer paso del escepticismo acerca de otras mentes. Nuestro autor esgrime con este fin la realización física de los *qualia* porque para él es un buen antídoto contra ese escepticismo, dejando de nuevo de lado la endiablada complejidad que se deriva del hecho de que la realizabilidad es en todo caso múltiple⁶. Y una vez despejado el camino de este modo no muy convincente, se pasará a establecer la posibilidad lógica de la inversión intersubjetiva del espectro a partir de la evidencia que hablaría a

⁶ Johnsen (1986) ha criticado con especial acritud que nuestro autor a veces hasta parezca defender una “realización física uniforme” de los *qualia*.

favor de la inversión intrasubjetiva. Es el *intra-inner argument* de Shoemaker, un argumento que no deja de complacerse en desenmascarar la incoherencia de Wittgenstein al aceptar la una y rechazar la otra. Porque está claro que si podemos suponer que Arturo sufre inversión intrasubjetiva en el tiempo *t*, cosa que no ocurre a otros, entonces antes o después de *t* (o ambas cosas) la experiencia cromática de Arturo tiene que haber sido totalmente diferente de la de las otras personas. Pero si admitimos que puede darse inversión intersubjetiva en los casos de inversión intrasubjetiva tenemos que admitir también que podría darse inversión intersubjetiva sin inversión intrasubjetiva. “Ahora bien, si esto es una posibilidad, entonces parece perfectamente coherente suponer, y perfectamente compatible con toda la evidencia conductual que tenemos sobre las experiencias de los otros, que, para ponerlo en las palabras de Wittgenstein, *una mitad de la humanidad tiene una sensación de rojo y la otra mitad otra*” (329). Es bien cierto, reconoce nuestro autor por otra parte, que la estructura misma de nuestro espacio cualitativo cromático podría hacer empíricamente imposible la inversión del espectro⁷. Pero al filósofo le basta con imaginar criaturas como nosotros pero con los espacios cualitativos del color simétricos, y careciendo de asociaciones emocionales innatas. Para ellas la inversión intersubjetiva sí que sería una posibilidad.

Una vez más: la mera posibilidad de inversión cualitativa intersubjetiva nos hace caer en la cuenta de que habría un aspecto crucial de nuestra experiencia (qué siente uno al tenerla) que esquiva las caracterizaciones funcionales. Pero para Shoemaker *que se dé* este aspecto es en sí mismo un hecho funcional, por tanto compatible con el funcionalismo y el fisicalismo. En otras palabras, la posibilidad de la inversión cualitativa no atenta contra el compatibilismo porque, aunque muestra que hay un sentido en que efectivamente los *qualia* no son definibles funcionalmente, en nada excluye que sus condiciones de identidad sí lo sean. Y, como insinuamos, la realizabilidad física de las propiedades funcionales abre la puerta a la posibilidad de descubrir las semejanzas y diferencias entre distintos sujetos, es decir, a la oportunidad de apresar empíricamente la inversión cualitativa. En definitiva, en adelante habrán de andarse con cuidado los que utilizan la posibilidad del espectro invertido para quebrar el compatibilismo. Porque a fin de cuentas podemos imaginar la inversión intersubjetiva porque podemos

⁷ Después de todo, no hay una proyección de matices de color en matices de color diferentes que proyecte gradaciones de color únicas –rojo puro, amarillo puro, verde puro y azul puro– en gradaciones de color únicas, y gradaciones binarias en otras gradaciones binarias, es decir, hay menos matices discriminables entre las gradaciones únicas verde puro y azul puro que entre las gradaciones únicas rojo puro y amarillo puro; además, rojo, naranja y amarillo se perciben, al parecer de forma no aprendida, como colores cálidos y positivos, mientras que verde y azul como fríos y negativos (1996, 150).

imaginar perfectamente la intrasubjetiva, y no se negará que las semejanzas y diferencias cualitativas que son la sustancia de esta última son funcionalmente definibles.

V

Pero donde la objeción de los *qualia* se ha mostrado más devastadora en sus efectos contra el programa funcionalista ha sido en los *thoughtexperiments* de los *qualia* ausentes, del tipo del de la cabeza homuncular o del de la nación china, casos en los que Block y Fodor vieron la radicalización del espectro invertido. Por todo lo que sabemos, así se plantea de ordinario el asunto, sería posible que dos estados psicológicos fuesen funcionalmente equivalentes aunque sólo uno de ellos tuviera contenido fenomenológico. Me duele pero no siente nada: el funcionalismo es entonces *evidentemente* insostenible, por mucho que algunos piensen que esto no es otra cosa que pedir la cuestión.

La Tesis de los *Qualia* Ausentes puede entenderse como si quisiera decir que los estados cualitativos no son funcionalmente definibles *en el sentido débil*, o bien como si significara que no lo son en el sentido *fuerte*. La primera versión implica la segunda, está claro, pero la segunda no implica la primera. El argumento de Shoemaker contra la posibilidad lógica de los *qualia* ausentes, que sin duda viene exigido por el mismo núcleo de su posición compatibilista, se dirigiría en concreto contra TQA-I. Y nosotros vamos a servirnos de la meridiana reconstrucción de este argumento que llevó a cabo Doore (1981, 391-392):

- (1) Supongamos que los *qualia* ausentes son posibles: S_n y S_q son estados funcionalmente equivalentes, pero S_q tiene asociados *pain qualia* y S_n no.
- (2) Hallarse en S_q da lugar, típicamente, a una creencia cualitativa, la de que uno está experimentando *pain qualia*.
- (3) Entonces, como S_q es funcionalmente equivalente a S_n , estar en S_n dará también lugar a la creencia cualitativa de que uno está experimentando *pain qualia*.
- (4) Entonces, un sistema en S_n podría creer que tiene *pain qualia* cuando no los tiene.
- (5) Si nosotros pudiésemos creer que tenemos *pain qualia* cuando no los tenemos, entonces ninguna evidencia de ninguna clase (conductista, introspectiva, fisiológica) distinguiría a los casos de *qualia* ausentes de los de *qualia* presentes.
- (6) Si pudiera haber un rasgo f de algún estado mental, tal que su presencia o ausencia no afectase la tendencia de ese estado a efectuar otros estados, conductas, etc., entonces ese rasgo sería totalmente incognoscible.
- (7) Así que, si los *qualia* ausentes son posibles, los *pain qualia* serían totalmente incognoscibles, necesariamente inaccesibles a la introspección.
- (8) Pero los *pain qualia* son accesibles a la introspección.

- (9) Por lo tanto, la suposición de que los *qualia* ausentes son posibles es falsa.

Block (1980) debilita la Tesis de los *Qualia* Ausentes para poder mantener, contra nuestro autor, su posibilidad lógica. No se trataría entonces de que el dolor pudiese carecer de carácter cualitativo y *sin embargo tener los mismos efectos* que el dolor “genuino”, sino simplemente de que podría haber un ser que tuviese *ersatz pain*, estado funcionalmente idéntico al dolor pero sin contenido fenoménico, aunque el dolor genuino seguiría siendo crucial en orden a producir las consecuencias que produce *de otro modo* el *ersatz pain*. Por mucho que estuviéramos en condiciones de asegurar que esta propuesta blockiana es del todo inteligible, lo que sin duda sería mucho admitir, quedaría el problema de que atacar con efectividad al funcionalismo requiere el sentido pleno de la TQA, o sea, la suposición de que el *ersatz pain* tiene las mismas consecuencias, y no de otro modo, que el dolor genuino.

De manera semejante Conee (1985) se figura desbaratar el argumento de Shoemaker con la afirmación de que la presencia del carácter cualitativo del dolor *makes a difference* en sus capacidades causales, diferencia que sí nos permitiría distinguir los dolores genuinos de los *ersatz*: el genuino, por poner un ejemplo, generaría un ítem de conocimiento distinto del que genera el *ersatz*. Pero tenemos que decir que por nuestra parte seguimos sin ver en qué puede afectar todo esto al funcionalismo. Tampoco entraremos a discutir ensayos eliminacionistas como el de Averill (1990), encaminados todos ellos en general a convencernos de que tomar la decisión de que no hay nada parecido a los dolores genuinos reportaría enormes ventajas conceptuales.

Y es que hay otro problema más serio, en el que entraría por ejemplo el ya citado Doore: ¿cómo hay que entender la *qualitative belief* en que en tan grande medida reposa el argumento shoemakeriano?. ¿Se trata simplemente de mi creencia de que “yo estoy teniendo *esta experiencia*”, donde “esta experiencia” refiere a la sensación que estoy entonces teniendo?. Porque en ese caso sería necesariamente verdadero que al tener una creencia cualitativa yo estoy teniendo esta o aquella sensación, de modo que los *qualia* ausentes serían imposibles por definición. Y podría parecernos que esto es así nada más que porque creer que se está teniendo una sensación y tenerla efectivamente son la misma cosa, con lo que está claro que el argumento de nuestro pensador deja al punto de funcionar. La conciencia de los *qualia* y la creencia cualitativa sería lo mismo, lo que resultaría letal para Shoemaker, a no ser que hagamos de los *qualia* otra cosa, a saber, nociones teóricas que es necesario introducir para dar sentido a las experiencias, y en particular a las cualidades secundarias. Pues bien, a esto precisamente parece llegar nuestro autor: “hay un sentido en el que no somos conscientes de *qualia* ni de similitudes fenoménicas” (1991/1996, 138). Con lo que se pondría otra vez de manifiesto que el compatibilismo se sigue viendo en dificultades si no restringe su ámbito para eludir el problema de la conciencia.

BIBLIOGRAFÍA

- Averill, Edward W.: "Functionalism, The Absent Qualia Objection and Eliminativism", *The Southern Journal of Philosophy*, XXVIII, No. 4 (1990): 449-467.
- Block, Ned J.: "Are Absent Qualia Impossible?", *The Philosophical Review*, LXXXIX, No. 2 (April 1980): 257-274.
- Block, Ned J. y Fodor, Jerry A.: "What Psychological States Are Not", *The Philosophical Review*, LXXXI, No. 2 (April 1972): 159-181.
- Conee, Earl: "The Possibility of Absent Qualia", *The Philosophical Review*, XCIV, No. 3 (July 1985): 345-366.
- Doore, Gary: "Functionalism and Absent Qualia", *Australasian Journal of Philosophy*, LIX, No. 4 (December 1981): 387-402.
- Johnsen, Bredo C.: "The Inverted Spectrum", *Australasian Journal of Philosophy*, LXIV, No. 4 (December 1986), 471-476.
- Putnam, Hilary: *Reason, Truth and History*, Cambridge University Press, 1981.
- Shoemaker, Sydney: "Functionalism and Qualia" (1975), *Identity, Cause and Mind. Philosophical Essays*, Cambridge University Press, 1984, 184-206.
- _____: "Phenomenal Similarity" (1975), *Op. Cit.*, 159-184.
- _____: "Some Varieties of Functionalism" (1981), *Op. Cit.*, 261-287.
- _____: "The Inverted Spectrum" (1981), *Op. Cit.*, 327-358.
- _____: "Absent Qualia are Impossible-a Reply to Block" (1981), *Op. Cit.*, 309-327.
- _____: "Qualities and Qualia: What's in the Mind?" (1990), *The First Person Perspective and Other Essays*, Cambridge University Press, New York, 1996, 97-121.
- _____: "Qualia and Consciousness" (1991), *Op. Cit.*, 121-141.
- _____: "Lovely and Suspect Ideas", *Philosophical and Phenomenological Research*, LIII, No. 4 (December 1993): 905-910.
- _____: "Phenomenal Character", *Noûs*, XXVIII, No. 1 (1994): 21-38.
- _____: "Self-Knowledge and 'Inner Sense' (Royce Lectures). Lecture III: The Phenomenal Character of Experience" (1994), *The First-Person Perspective and Other Essays*, 246-268.
- _____: "Intrasubjective/Intersubjective", *Op. Cit.*, 141-157.
- Tye, Michael: "The Subjective Qualities of Experience", *Mind*, XCV, 377 (January 1986): 1-17.
- White, Nicholas P.: "Prof. Shoemaker and So-called 'Qualia' of Experience", *Philosophical Studies*, 47 (1985): 369-383.

Mariano Rodríguez González.
Universidad Complutense de Madrid
e-mail: marian@psi.ucm.es



MENTAL CAUSATION: The supervenience argument and the proportionality constraint

JÜRGEN SCHRÖDER
University of Heidelberg

ABSTRACT: In the first part of this paper I argue that Kim's principle of causal inheritance which is supposed to secure causal relevance to second-order functional properties cannot be sustained. So functional properties, although physical, have no causal potential at all. Then I address the so called generalization problem. I'll argue that Kim's response to this problem is inadequate and that therefore the generalization problem continues to be a threat to the causal potential of any higher level property.

In the second part I take Stephen Yablo's proportionality constraint to be a solution for both the problem of mental causation proper and for the generalization problem. Contrary to Yablo, however, I argue that the proportionality constraint should be divorced from the relation between determinable and determinate properties.

The third part tries to compare content properties with intrinsic properties of mental representations. I argue that the semantic property of a representation is more proportional to an abstract behavioral property than its intrinsic properties.

KEYWORDS: Mental causation, causal inheritance, functional properties, proportionality, determinables/determinates, generalization problem.

Introduction

Over the last years Jaegwon Kim has argued that the possibility of mental causation is called into doubt by an argument which unfolds in three stages. The first stage consists of the principle of causal closure of the physical domain which says, roughly, that for every phenomenon describable in physical terms there is a sufficient cause equally describable in physical terms. When we apply this principle to any case of mental causation it says that for the event to be explained, e.g. a behaviour, there is a physical cause of this behaviour. Now we have a mental and a physical cause for the same behaviour and the question arises how they are related to each other. The

second stage of the argument tries to consider exhaustively the possible candidates for such relations yielding the result that the mental properties of the cause supervene on its physical properties without being identical or reducible to them. Finally an exclusion principle is applied which states that if the complete causal work on a given occasion with respect to a property B (of behaviour) is done by a property P then anything which is not P is causally superfluous with respect to B on this occasion. Since the result of the second stage is that mental properties only supervene on physical properties without being identical with them the conclusion follows that mental properties are causally idle.

First Part: Kim's premises

Kim has presented this argument not in order to convince us of epiphenomenalism but in order to state the background against which the possibility of mental causation has to be demonstrated. For him the reality of mental causation is not in question. Rather, what has to be done is to show how mental causation is possible in a physical world. If Kim's argument purports to show why mental causation is not possible and if mental causation is actual then there must be something wrong with this argument. There must be one or several places which should not stand up to closer inspection.

Closure, supervenience, and causal inheritance

I propose to accept the first stage of the argument, the principle of causal closure. The main reason for accepting it is the past track record of the physical sciences such as physics proper, physical chemistry, chemistry, biochemistry etc. which, in order to causally explain any phenomenon in their domain, needed to make no reference to supernatural powers.

When we accept the closure principle vis-à-vis semantic properties there are two stages left where something might have gone wrong. The exclusion of all other relations except supervenience and the exclusion principle itself. I suggest to accept the second stage of Kim's argument since if a kind of dependence relation is to obtain between mental and physical properties mental properties will not be independently instantiable and so the possibilities of overdetermination, of joint sufficiency and of intermediate links in a chain are ruled out. For semantic properties identity and realization seem to be ruled out as well. So mental and intrinsic physical properties seem to be distinct properties (no identity), the one not being realized by the other

(semantic properties are not functional roles) but the mental depending on the physical.

There is perhaps one point which is controversial in this part of the argument. It is the claim that semantic properties are not functional roles. I have deliberately put this claim as an identity thesis because I think that in this form it is less controversial than as a determination thesis. One may plausibly hold that the content of a brain state is partly determined by its relation to other brain states and partly by its relation to a certain condition in the world (outside the brain), but to claim that content is identical with the functional role of some physical property of a brain state is clearly misconceiving the fundamental character of intentional states, viz. that they are directed at objects. A functional role, on the other hand, is not directed at anything.

Assume for a moment, however, that semantic properties *were* functional roles, that they stood in a realization relation with intrinsic properties of brain states. In this case, according to Kim, the possibility of mental causation would not be in question anymore. This is because as functional properties they would inherit their causal powers from their realizers.

The principle of causal inheritance according to which a functional property inherits the causal powers of its realizer seems to be wrong, however, for the following reason. If the metaphor of inheritance is to be any guide in these abstract matters there must be a heir and a testator. But we only have a testator, the physical property, and no heir. This is because the functional property is second-order. It is not a property on the same footing as the realizer property. Just as a policeman cannot inherit the fortune of Jones if Jones *is* the policeman, i. e. if Jones fills the role of a policeman, so it is impossible for a second-order property to inherit the causal powers of its first order property. If there cannot be any inheritance where there is only a testator but no heir there is no inheritance of causal powers by functional properties because there is no heir.

The same point can be seen when we think about the second-order property of being a functional role of the first-order property. A functional role is a set of causal powers. The first-order property is said to have this set of causal powers. If the second order property would inherit the causal powers from the first-order property it would inherit itself since it *is* what it is supposed to inherit, viz. the set of causal powers. But again, if inheritance requires two distinct entities then we have no instance of inheritance here.

Exclusion

We have arrived at the third stage of the argument, the exclusion principle which says that since mental properties supervene on physical properties and are not identical with them they have no causal work to do because this work is already done by the physical properties. I think the basic mistake of the argument is hidden in this principle which at first sight looks completely innocent and self-evident. What could be more obvious than the fact that if somebody's work is done by somebody else then there is no need anymore for anyone to do this work? The agent who does the work has to bring about a certain event *and* he has to make sure that it satisfies the requisite property. He has to make sure that a possible variety of physically different events all satisfy the relevant description. What the physical property P_1 makes sure is that another event has a certain property P_2 . It does not and it cannot assure that a different property P_3 equally satisfies a certain description M^* . What M_1 seems to be able to make sure is precisely what P_1 is unable to do, namely to make sure¹ that some P_i of a set P of physical properties which satisfies M^* gets instantiated. What M_1 , the property of being an intention that the gas tank gets filled, does is to make sure that this goal is achieved and that some P_i , one of the members of the set of goal equivalent behaviours, gets instantiated. And it is clearly relevant for the achievement of the goal that if a certain P cannot be instantiated another P takes its place.

I conclude that the analogy of work being done is misleading in an important respect: the work being done, in the sense of making sure that if one P_i fails to bring about a certain behaviour another P_j does the job, *could* not be done by subvenient physical properties.

The generalization problem

One possible reaction to Kim's argument is to endorse all of its premises and its conclusion and to argue that the same kind of reasoning shows that all properties except those of the ultimate constituents of things are epiphenomenal. A variation on this theme is the assumption that there is no basic level of reality, no level of last constituents. In this case there would be no causation whatsoever (Baker, 1993; Van Gulick, 1992; Block, 1997).

¹ „Making sure“ should not be understood deterministically. If Jones makes sure that the car tank gets filled it is still possible that it does not get filled. The same goes for properties.

Kim tried to rebut the generalization thesis. The generalization argument assumes that the supervenient and subvenient properties belong to different levels of organization such that the supervenient-subvenient relation correlates perfectly well with the part-whole relation. Supervenient properties belong to wholes while subvenient properties belong to the parts of these wholes. Because of that assumption it could seem that the doubtfulness of the causal efficacy of mental properties transfers to their supervenience base if this base does itself supervene on further micro-properties.

But it is wrong, says Kim, that the supervenience relation tracks the macro-micro relation. And since the supervenient mental properties belong to the same particular, viz. the whole brain, to which the subvenient physical properties belong there is no iteration and no draining away of causal powers of these physical (physiological) properties. This is one half of Kim's response. It is designed to stop the regress of causal powers from one property at a more complex level to another property at a less complex level. The second half of his response is intended to show how macroproperties can be causally efficacious. To this purpose Kim introduces the notion of a micro-based property:

(MB) "P is a micro-based property just in case *P* is the property of having proper parts, $a_1, a_2, \dots, a_n$, such that $P_1(a_1), P_2(a_2), \dots, P_n(a_n)$, and $R(a_1, \dots, a_n)$." (Kim, 1997, p. 292)

He uses the notion of micro-based properties to delimit the domain of physical properties. A property is physical either if it is mentioned in physical theory or is a functional property of those or is micro-based on them.

Now, if *P* is a micro-based macroproperty it can have effects which none of the P_i have and in this sense exceeds the causal powers of all the P_i taken severally. This is how physical macroproperties can be causally efficacious, i. e. have causal powers not had by the microproperties. So it is false that every property mentioned by the special sciences is likewise in doubt concerning its causal efficacy as are the mental properties. The problem of mental causation does not generalize.

The examples Kim gives for micro-based properties are of two sorts. There is the case of being a H₂O molecule which is a structural property and fits Kim's definition (which is based on Armstrong (1978)) perfectly well and there is the case of mass which, I think, does not fit it at all. The property of mass is *not* the property of having proper parts a_1 - a_n , with properties $P_1(a_1)$ - $P_n(a_n)$ and relation *R*. The reason for this is simply that lots of things with a

different constitution can share the property of having a certain mass, 1 kg say. All these things like pieces of cork, gold and marmelade have a different molecular or atomic constitution and for each such thing the property of 1 kg would, of course, have to be analyzed in a different way. For this property none of the different analyses would be *the* correct analysis whereas there is one such analysis in the case of being a water molecule. I take the absence of such an analysis to suggest that mass is not a structural property at all although it is clearly micro-based in a sense, since a typical body of our common sense world would not have a mass of 1 kg if it had no parts with masses that sum up to 1 kg. Although it is micro-based in this sense it is not a structural property and is not covered by Kim's definition of micro-based properties.

In order to cover properties which don't imply anything with respect to the composition of the whole of which they are properties we might extend the definition somewhat.

(MB') A property of a whole *w* is micro-based iff its instantiation depends on the instantiation of properties of the parts of *w*.

According to this new definition virtually every property of a material object which has parts would be micro-based. But if everything which is a property of a whole is micro-based then surely mental properties are micro-based too if we assume that substance dualism is wrong and that mental properties belong to material substances, e. g. to nervous systems.

In order to sustain what I think is right in the generalization argument I want to concentrate on a certain difficulty in Kim's account of micro-based properties and how they can be causally efficacious. Consider the following quote:

"But this only means that the causal powers of the micro-based property *P* are *fixed*, or *determined*, by the causal powers of the properties and relations, *P*₁, ..., *P*_n, and *R*, that figure in *P*'s construction, but they need not be, and are not likely to be, *identical* with the causal powers of these constituent properties and relations. There is a world of difference between *determination* and *identity*. That is how micro-based properties can bring new causal powers into the world." (Kim, 1997, p. 296)

We have three things here. The macroproperty P of a whole w , the P_i which are properties of the parts of w , and the construction $R(P_1-P_n)$ ². What determines P is $R(P_1-P_n)$, not any of the P_i . And the fact that the relation between P and $R(P_1-P_n)$ is one of determination and not one of identity is taken by Kim to explain how micro-based properties can bring new causal powers into the world. Initially the causal powers of P were compared with the causal powers of the constituent properties P_i and it was claimed that those of P typically exceed those of these P_i . Now, however, the causal powers of P are compared with the causal powers of $R(P_1-P_n)$ and it is claimed that those of P exceed those of $R(P_1-P_n)$. This, at least, seems the most plausible interpretation of the quoted passage since the P_i taken severally don't determine P and since the difference between determination and identity is supposed to account for the extra causal powers of P .

If we accept this interpretation a contradiction is generated between this claim of how the causal powers of micro-based properties can exceed those of their basis and another principle which Kim endorses and which we could call the principle of non-excess of causal powers (NE).

(NE) If property M is instantiated on a given occasion in virtue of being realized by (or in virtue of its supervenience on) property N , then the causal powers of M are identical with (at least, cannot exceed) the causal powers of N . (Kim, 1993, p. 170)

If (NE) is correct P cannot have causal powers which $R(P_1-P_n)$ doesn't have because $R(P_1-P_n)$ determines P (i.e. P supervenes on $R(P_1-P_n)$) and because (NE) says that in cases of supervenience P has no additional causal powers with respect to its supervenience base. Even if it is true that P has causal powers which none of the P_i s have this fact does not seem to do much in order to counter the generalization argument. Since this argument could be recast in terms of supervenience bases $R(P_1-P_n)$ which determine the corresponding macroproperty P and which have all the causal powers that P is said to have. The generalization argument doesn't need to make reference to the part-whole hierarchy. It can be put forward for properties of one and the same level. If they stand in a supervenience relation and are not identical then the supervenient properties seem to be causally idle. So it seems that the problem of mental causation *does* generalize.

² For the sake of brevity I denote the complex supervenience base of the macroproperty by $R(P_1-P_n)$. This supervenience base consists of the properties of the parts $P_1(a_1)-P_n(a_n)$ and the relation R between the parts a_1-a_n .

I think we should not take the generalization argument lightly but try to find an account which does not only solve the problem of mental causation but also the problem of how micro-based physical properties can be causally efficacious. In the next section I will consider Stephen Yablo's proposal with respect to its potential for answering both questions

Second part: Yablo's proportionality constraint

The idea that a property makes sure that an effect occurs which satisfies a certain description meshes well with a constraint on causation which has been suggested by Stephen Yablo: the proportionality constraint (Yablo, 1992, 1997).

(PC) For any instance of causation the property of the cause which is to be causally relevant for the instantiation of a given property of the effect must be proportional to this property of the effect.

The proportionality constraint works in two directions. On the one hand, and more obviously, it forces explanations to deliver the details which are needed to account for the details of the effect. On the other hand, less obviously, it asks for the maximally abstract property which is able to account for the property of the effect to be explained. If we accept (PC) as a constraint on causation in general then we are in a position to see what is wrong with the exclusion principle. Instead of exempting the supervenient mental properties of their causal efficacy the subvenient properties which contain more detail than the supervenient ones are causally irrelevant with respect to certain properties of the behaviour to be explained. If the property to be explained is one of drinking then the exact form of the physical state on which the intention to drink supervenes is irrelevant for this property to be instantiated. If, on the other hand, the exact trajectories of the limbs, which realize a drinking on a given occasion, are to be explained then to be an intention to drink contains not enough detail and the form of a physical state in the motor cortex may be proportional to this property of the effect.

The proportionality constraint counteracting the force of the intuition which sustains commitment to the exclusion principle, does equally show how properties which are mentioned by other special sciences such as biochemistry, physiology, geology, etc. can be causally efficacious. They can be causally efficacious as proportional properties. Only if they are purely

functional properties³ they cannot have causal powers because then they *are* causal powers. The causal efficacy of the properties singled out and mentioned by the special sciences is even *guaranteed* since they have been accepted by the practitioners of those sciences precisely *because* they are proportional to the effects to be explained. So (PC) seems not only to lay to rest worries about mental causation but also worries that were inspired by the generalization argument.

Proportionality and the relation between determinable and determinate properties

I have introduced proportionality in terms of corresponding degrees of abstractness. Since proportionality is characterized with respect to missing or superfluous detail it is natural to suppose that those properties which contain more detail and those which contain less stand in a determinate-determinable relation to each other. This is in fact Yablo's proposal. In his account of mental causation the proportionality constraint is run together with the determinable-determinate relation. If mental properties are determinable vis-à-vis physical properties then they could be proportional to an abstract behavioural property whereas the properties of the brain state which are more determinate would not be proportional and thus the causal efficacy of mental properties would be secured.

Although it is natural to combine the proportionality constraint with the determinable-determinate relation it is not necessary. The absence of a certain kind of detail from a property of the cause does not entail that this property is a determinable of another property of the cause. Proportionality is a relation between a property of the cause and a property of the effect whereas the determinable-determinate relation is supposed to hold between properties of the cause. If the detail in the proportional property of the cause is missing that doesn't mean that the property is a determinable of another property which contains just this detail. It only means that the detail is missing, period. Inferences as to the relation between the proportional property and other properties of the cause are not warranted by the proportionality relation.

³ A property is *purely* functional iff it is the second-order property of being a set of causal powers which is a property of a first-order property. It is, of course, possible to refer to a first-order property via a functional profile. In this case the property referred to is not functional and thus *can* have causal powers.

And it had better be this way because Yablo's idea that mental properties are determinable relative to physical properties suffers from a serious problem.

The problem for this proposal is that we don't see how *semantic* properties (whose causal efficacy is at stake) *could* be determinables of physical properties of the brain. To be crimson is to be red in a certain way. The same is not true of physiological brain properties. To be an activation vector of form F or to be a set of action potentials with characteristics C does in no way have as its core the representation of something.

On the contrary, it seems that the semantic properties of a representation are rather unconstrained by its physical properties.

Since this problem suggests that semantic properties are not determinables of physical brain properties I propose to separate the proportionality constraint from the determinable-determinate relation. Semantic properties of brain states can be proportional to abstract properties of behaviour without being determinables of physiological properties.

Third part: Applying the proportionality constraint to semantic properties

Content and activation vectors

In order to show now that the semantic properties are more proportional to the relevant property of behaviour it is first of all necessary to say what the properties are on which the semantic properties depend. Let us assume that the base properties are the forms of activation vectors where an activation vector is a n-tuple of activation values of a set of nerve cells (McClelland et al., 1986, Churchland & Sejnowski, 1992, Georgopoulos, 1997).

Let's assume then that the forms of activation vectors are the base properties for content. And let's counterfactually assume for a moment that every vector consisted of the activation of a single cell. Every representational content would be represented by single cells then. Assume further, again contrary to fact, that all these cells had the same intrinsic characteristics, i.e. the same input-output functions, the same thresholds, etc. Now, consider one of these cells in the brain of a cat which, say, represents mice, is connected to sensory representations of mice and to motor representations of chasing behaviour. Another cell represents dogs, is connected to sensory representations of dogs and to motor representations of fleeing behaviour. In

this case we could substitute one cell by the other without there being any change either in consciousness or in content or in behaviour. If we want to explain chasing behaviour we refer to what the cell's activation represents, not to the physical properties of the representation. If the cell had not represented mice but dogs it would have activated the fleeing behaviour motor assembly. It would have represented dogs if it had been connected to sensory representations of dogs such that its maximal response would have been to dogs and not to mice. In this case the base properties do not contain too much detail but they are not specific enough in order to explain the differences in behaviour since we have assumed that they are intrinsically all the same.

Let's assume now that every cell differed intrinsically from every other cell and had a different content. Would the intrinsic properties then preempt the content properties from being causally efficacious? That depends on how much the intrinsic properties of a given cell can vary without affecting its role of being a mediator between sensory and motor representations. In any case, if the thresholds and rules are described arithmetically slight variations of the numbers involved should not affect that role. But if it doesn't then these properties are not proportional to the behaviour to be explained. Recall now that these behavioural properties are not described in terms of limb trajectories but quite abstractly. So even if the kind of movement covered by a trajectory description were different this would not mean that the behaviour relative to a more abstract description such as "chasing", "hiding", "attacking", etc. would be different. In fact, if the description is not factive, i.e. if it can be satisfied without a certain goal being reached, then the probability seems very high that a slightly different movement would count as a token of the same type of behaviour. But if that is the case certain variations on intrinsic properties don't lead to different behaviour. And if they don't they are not proportional to the type of behaviour in question.

Until now we have considered representations being constituted by single cells. If our mental representations were activations of single cells their intrinsic properties would not be proportional to the properties to be explained. But initially and more realistically we assumed the bases of content properties to be the forms of activation vectors. Does this difference make a difference for our discussion? I think that, basically, the situation is the same whether you think about it in terms of single cells or in terms of activation vectors over cell assemblies. Assume that two activation vectors are intrinsically the same but different with respect to their connection with

sensory representations and what has been said about single cells still applies. Likewise if they are different but their difference does not exceed certain limits the caused behaviour will still belong to the same abstract type thereby showing the forms of activation vectors to be unproportional to the effect to be explained. We may conclude that whether mental representations are constituted by single cells or cell assemblies is a difference which makes no difference with respect to the causal efficacy of content properties.

Proportionality and exclusion

We have tacitly assumed in this discussion that the content of a representation is more proportional to the property of behaviour to be explained than any intrinsic physical or physiological property. But this assumption may be wrong. Perhaps there are abstract physical properties which are just as proportional to the behavioural property as the content of a representation is. And if there are we could run a variant of Kim's exclusion argument in order to show that content properties are causally idle since we already have a proportional physical property and since this property is not identical with the semantic property. How are we to respond to this objection?

Even if we had possible candidates for such abstract properties the problem that counted against Yablo's idea that semantic properties are determinables of physical ones now counts against a modified version of Kim's exclusion argument. The physical properties of two representations may be the same whereas their meanings are different. In every such case the physical property would be too unspecific in order to explain the difference between two behaviours associated with the two representations. That is, it would not meet the proportionality constraint. The situation is that the semantic properties are proportional to the behavioural properties in all cases whereas the physical properties are *not* proportional in all cases. Now, given this relation between physical and semantic properties with respect to proportionality how *could* the (abstract) physical properties pre-empt the semantic ones? For a property P_1 to pre-empt another property P_2 causally it has to cause another property P_3 in the first place. But if it's not proportional to P_3 it doesn't cause it. If there is really any pre-emption with respect to abstract behavioural properties then it seems that the semantic properties, in virtue of their proportionality, pre-empt the physical properties rather than the other way 'round.

REFERENCES

- Armstrong, D., *A Theory of Universals*, Vol. II, Cambridge: Cambridge University Press, 1978.
- Baker, L. R., "Metaphysics and Mental Causation", in J. Heil and A. Mele (eds.), *Mental Causation*, Oxford: Clarendon Press, 1993, 75-95.
- Block, N., mentioned in Kim, 1997.
- Churchland, P. & Sejnowski, T., *The Computational Brain*, Cambridge, Mass., MIT Press.
- Georgopoulos, A., „Voluntary movement: computational principles and neural mechanisms“ in M. D. Rugg (ed.), *Cognitive Neuroscience*, Hove: Psychology Press, 1997, 131-168.
- Kim, J., „Mental Causation in a Physical World“, *Philosophical Issues*, 3, 1993, 157-176.
- Kim, J., Does the problem of mental causation generalize?, *Proceedings of the Aristotelian Society*, 97, 1997, 281-297.
- McClelland, J. L., Rumelhart, D. E., and the PDP Research Group, *Parallel Distributed Processing*, Vols. I and II, Cambridge, Mass.: MIT Press, 1986.
- Van Gulick, R., "Three Bad Arguments for Intentional Property Epiphenomenalism", *Erkenntnis*, 36, 1992, 311-31.
- Yablo, S., "Mental Causation", *Philosophical Review*, 101, 1992, 245-80.
- Yablo, S., "Wide Causation", *Philosophical Perspectives*, 11, 1997, 251-81.

Jürgen Schröder

University of Heidelberg

e-mail: jschroel@urz-mail.urz.uni-heidelberg.de



Practical Philosophy

Filosofía Práctica





LOS DERECHOS POSITIVOS DESDE UNA PERSPECTIVA ANALÍTICA

F. JOSÉ AUSÍN

CALIJ

LORENZO PEÑA

Instituto de Filosofía CSIC

ABSTRACT: In this paper we provide a twofold analysis of positive rights, logical and juridical. We use deontic logic as the appropriate tool in order to clarify the concept of positive right. In this way, an adequate treatment of quantification becomes absolutely necessary since positive rights concern with some goods or others. Then, we give two important proofs: the principle of ensuant obligations for positive rights, and the principle of legitimate claims. Finally, we conclude with the requirement of taking positive rights seriously, because it would be a mockery to promulgate positive rights and yet to leave them empty of contents.

KEYWORDS: Positive rights, deontic logic, quantification, political commitments.

Derechos

Tener un derecho a algo es que a uno le sea lícito tener o hacer ese algo. Es lo que se denomina como “derecho subjetivo” que, siguiendo a Francisco Suárez, puede definirse igualmente como la capacidad o posibilidad de hacer o exigir alguna cosa por considerarse legítimo o estar así establecido. Algunos han defendido una distinción entre derechos consistentes en mera licitud o autorización y derechos en un sentido más fuerte, en el sentido en que los demás tengan la obligación de no impedir o estorbar el ejercicio de tal derecho (principio de no impedimento, al que volveremos más adelante). Así se habla de permisos o derechos débiles, en el sentido en que se refieren a cursos de acción lícitos por el hecho de no estar prohibidos, y de permisos fuertes, en cuanto explícitamente reconocidos en ordenamiento legal¹. A nuestro entender, si el anterior principio de no impedimento no tuviera

¹ Para seguir la polémica en torno a la distinción entre permisos fuertes y permisos débiles, véase Alchourrón y Bulygin (1984[1991]).

vigencia para los permisos débiles, una sociedad tal no garantizaría a sus miembros la capacidad de hacer lo no prohibido y la de abstenerse de hacer lo no obligatorio, echando por tierra otro principio básico de las relaciones entre normas: a saber, que lo no prohibido está permitido (principio de permisión). Se darían entonces situaciones en la que se autorizara impedir hacer una determinada acción a un individuo, sin que éste tuviera prohibido hacer tal acción; o, sin haber obligado al individuo hacer esto o aquello, permitiría a otros que lo fueren a hacerlo, a hacer algo tal que le es lícito no hacerlo.

Derechos positivos

Veamos ahora a qué llamamos ‘derechos positivos’. No debe confundirse este término con ‘derecho positivo’, que es el conjunto de leyes, normas, reglamentos, etc. explícitamente promulgados por el Estado.

Desde un punto de vista lógico-sintáctico, podríamos definir tales derechos positivos como aquellos que se formulan con enunciados cuyo *dictum* -aquello expresado después de la calificación normativa “está permitido que”- viene encabezado por un cuantificador existencial: El derecho a que haya *algo* así o asá (p.e. el derecho a que haya una vivienda a la cual tenga uno acceso, o a que haya alguna nacionalidad que uno posea, etc.); lo cual suele expresarse en el lenguaje corriente como el derecho a tener (o recibir, o adquirir, etc.) un (o una) X.

Sin embargo, esa pauta puramente gramatical no siempre es un buen hilo conductor, ya que el artículo indeterminado viene a menudo elidido, sin que se aprecie en la superficie de la enunciación ninguna señal clara y obvia de que se trata de un derecho positivo en ese sentido lógico-sintáctico. (Véanse, por ejemplo, los artículos del Capítulo tercero del Título I de la Constitución española de 1978, referidos a la política social y económica).

Otra caracterización de los derechos positivos es jurídica y de contenido. Se denominan ‘derechos negativos’ a aquellos que prescriben la ausencia de trabas o imposiciones para realizar una acción (‘libertades de’): El derecho a manifestar las opiniones (libertad de expresión), a asociarse con otros (libertad de asociación), a no pensar de un modo determinado (libertad de conciencia), etc., propios de las primeras Declaraciones de Derechos Humanos de inspiración liberal. En cambio, los derechos positivos son derechos que conllevan deberes no puramente negativos de los demás, sino deberes positivos; es decir, un derecho es positivo cuando su respeto requiere acciones y no meramente omisiones, cuando exige una determinada prestación. No hacer nada puede impedir a otro que consiga bienes a los que tiene derecho; por ejemplo, si existe carencia de viviendas para que los sin-techo encuentren cobijo y no se articula una política fiscal que aporte fondos públicos para la construcción de viviendas. Ejemplos similares pueden darse en relación a otras necesidades básicas reconocidas como derechos, como la atención sanitaria, la alimentación, la educación, etc. Por tanto, a diferencia de

los derechos negativos del liberalismo decimonónico, los derechos positivos no son derechos a que no haya nada así o asá, sino a que haya *algo* con ciertos rasgos; y por ende no exigen, para ser respetados, meramente que los demás no nos hagan nada (o nada de tal índole), sino que (por lo menos algunos de ellos) sí nos hagan algo. Respetar los derechos positivos conlleva una cierta prestación, recaiga en quien recayere la carga de tal prestación (el resto de los individuos, el Estado o la comunidad internacional).

Evolución y tipos de derechos positivos

Así entendidos, en cualquiera de las caracterizaciones que hemos dado, muchos de los derechos positivos son derechos económicos, sociales y culturales. La Declaración universal de los derechos humanos de 1948 ya apunta en la dirección del reconocimiento de los derechos positivos en este sentido: art. 23: derecho a trabajar; art. 25: derecho a un nivel de vida suficiente para el bienestar; art. 26: derecho a la educación; art. 28: derecho a que reine un orden socio-económico que haga posible y viable la satisfacción de tales derechos y el logro de tales bienes. Más explícito es el Pacto internacional sobre los derechos económicos, sociales y culturales de 1966, cuyo artículo 7 prescribe el derecho a disfrutar de condiciones de trabajo que conlleven remuneración equitativa y unos medios de existencia decentes; art. 9: derecho a la seguridad social; art. 10: derecho de las familias, las madres, los niños y jóvenes a la protección social; art. 11: derecho a la calidad de vida suficiente, vestido, alimentación, alojamiento y mejora continuada en el nivel de vida; derecho al reparto equitativo de los recursos alimenticios.

Sin embargo, aunque el reconocimiento de los derechos positivos es más bien reciente, no podemos soslayar que en el pensamiento jusfilosófico cristiano de la Edad Media y del Renacimiento -desde Tomás de Aquino a Campanella- no está ausente la idea de los derechos positivos. No obstante, entre que tales tomas de posición quedaban anegadas en un mar de salvedades, que sus autores no dedicaban en general esfuerzos denodados a predicar su aplicación, y que, en cualquier caso, las doctrinas liberales de la época barroca vinieron a desacreditar la vieja idea del bien común², el hecho es que poco se aplicó de todo eso. Pero aun dentro del apogeo del liberalismo filosófico y político brotará una tendencia de reconocimiento de los derechos positivos: La Constitución de la República francesa de 1793 estipula que la sociedad tiene el deber de socorrer a los ciudadanos en desgracia, sea dándoles trabajo, sea asegurándoles medios de vida cuando no puedan trabajar. Este es el remoto origen del ingreso mínimo de integración que en este final de siglo se ha empezado a establecer en algunos países desarrollados, como España, aunque todavía de un modo muy limitado y exiguo.

² Para una defensa del bien común como precepto básico, véase Peña (1997).

Por lo que respecta a nuestro país, ya en la Constitución de 1812, en su artículo 366, se establecía un derecho positivo: el de recibir educación. Pero no será hasta 119 años después, con la Constitución de la II República de 1931, cuando reaparezcan ciertos derechos positivos: Capítulo segundo del Título III (Familia, economía y cultura). -Con relación a los derechos económicos, sociales y culturales en la actual Constitución española de 1978 volveremos más adelante.

Mas, a pesar de lo dicho hasta aquí, hay que reconocer que ni todos los derechos sociales, económicos y culturales son positivos, ni todos los derechos positivos pertenecen a la esfera socio-económica y cultural. Así, derechos socio-económicos como el derecho de huelga, el de sindicación, el de exclusión de ciertas cláusulas en los contratos de trabajo, etc. difícilmente pueden verse como derechos positivos. Hay, desde luego, muchos casos dudosos. Por ejemplo, ciertos derechos laborales pueden interpretarse ya como derechos positivos (el derecho del trabajador a que se inserte alguna cláusula de cierto tenor en el contrato laboral), ya como negativos (a que no se inserte ninguna cláusula del tenor opuesto).

Por otro lado, hay derechos procesales que son implícitamente derechos positivos, como el derecho de los acusados de un delito a un proceso sin excesivas dilaciones y con garantías. Sólo pueden respetarse tales derechos procesales si el Estado dota suficientemente a la administración de justicia de los medios indispensables para esos cometidos. Incluso el derecho a ser juzgado por un jurado es un derecho positivo, cuyo respeto también acarrea acciones (y no meramente omisiones) por parte de los poderes públicos.

Igualmente, en ámbitos del derecho privado, como en el derecho civil, una serie de actos particulares crean derechos positivos en otros particulares: donaciones, herencias, legados, contratos varios.

Más aún, la penalización de la omisión de socorro conllevó el reconocimiento del derecho de cada uno a recibir socorro cuando concurren ciertas condiciones y, por consiguiente, la obligación derivada para los demás de ofrecer ese socorro cuando se den esas circunstancias (p.e. que se haga sin grave riesgo para uno o los suyos o para otros bienes jurídicamente protegidos, en proporción a la importancia del socorro y al esfuerzo requerido para darlo). Todos éstos últimos son un círculo de derechos y deberes que no suele incluirse en la esfera de lo socio-económico-cultural pero que, como vemos, envuelven acciones y prestaciones y no meramente omisiones, como en el caso de los derechos negativos.

En cualquier caso, es en el dominio de los derechos económicos, sociales y culturales donde los derechos positivos exigen un mayor *grado* de

prestaciones (acciones) y donde más problemático se vuelve su tratamiento adecuado³.

Lógica deóntica

Para analizar en detalle los derechos positivos y toda la suerte de implicaciones que conllevan se hace necesario recurrir a la lógica deóntica como instrumento de estudio de las conexiones que vinculan unos derechos con otros y con sus correspondientes deberes. De forma genérica puede decirse que la lógica deóntica pretende analizar las relaciones formales que se establecen entre obligaciones, permisos y prohibiciones:

Si es una obligación el votar por cualquier partido político, entonces ¿está permitido hacerlo? Si tengo derecho a circular por la comunidad europea, ¿está prohibido que me lo impidan?, ¿es un deber que se faciliten los medios para tales desplazamientos? Si estoy sometido a la obligación disyuntiva de hacer esto o lo otro, ¿cómo cumpliré ese deber, satisfaciendo lo primero, o lo segundo, o ambas prescripciones? Si tengo la obligación de asistir a los actos institucionales de mi universidad e igualmente tengo el deber de no perder ninguna de las clases de doctorado, ¿estoy sujeto a la obligación conjunta de ambas acciones? ¿Son las consecuencias causales de una acción permitida igualmente lícitas? Etc.

Éstas y muchas otras son preguntas que nos asaltan a la hora de reflexionar sobre el entramado de calificaciones deónticas (permisos, deberes, prohibiciones) que atraviesan todas las relaciones humanas. En definitiva, las expresiones que incluyen calificaciones de este tipo no son sino normas (jurídicas, morales, técnicas) cuya estructura y relaciones inferenciales trata de clarificar la lógica deóntica -también denominada por ello lógica de las normas. Es decir, la lógica deóntica es la teoría de las inferencias normativas válidas, es el análisis de las condiciones y reglas en que un razonamiento que incluya las calificaciones de prohibición, deber o permiso resulta correcto⁴.

La lógica deóntica que vamos a considerar en este trabajo está alejada de los habituales modelos calcados de la lógica modal, tomando sus principios básicos de la práctica misma de la vida legislativa y judicial. Esto supone el abandono de una serie de principios clásicos de la lógica deóntica estándar que estaban en el origen de las múltiples paradojas que presentan estos sistemas clásicos⁵. Además, el cálculo cuantificacional subyacente es la lógica transitiva, una extensión no conservativa de la lógica de la relevancia E, desarrollada por uno de los autores⁶. [Las notaciones convencionales son O

³ Sobre la caracterización y fundamentación de los derechos sociales resulta interesante el debate entre De Castro (1995) y Contreras (1995).

⁴ Sánchez-Mazas (1973, pp. 25 ss.).

⁵ [Díez] Ausín y Peña (1993).

⁶ Peña (1993).

para el operador de obligatoriedad, P para el de permisión o licitud y V (vedado) para el de prohibición].

Asumimos como básicos la equivalencia entre la obligatoriedad de p y la ilicitud de no p [$Op \equiv NPNp$], y el llamado principio de Bentham (formulado ya por Leibniz) por el cual lo que es obligatorio es también lícito [$Op \rightarrow Pp$]. No tiene sentido un sistema normativo que prescriba una obligación sin también garantizar el permiso para ejecutar la acción que está obligado a realizar.

Otro principio básico de la lógica deóntica es aquel por el que las consecuencias causales de una acción lícita son igualmente lícitas. Una formulación alternativa de este ‘principio de cierre causal’ es la siguiente: Las acciones que causan un resultado prohibido están prohibidas [$p \odot q \ \& \ \forall q \rightarrow Vp$]. De ahí que la instigación al crimen sea también un delito.

Pero sin duda el principio más importante de la lógica deóntica objetiva -un principio ampliamente aceptado por los filósofos del derecho- es el ya mencionado ‘principio de no interferencia’⁷: Si alguien tiene derecho a hacer o tener algo, es ilícito que los demás le estoben e impidan realizar o tener ese algo. Y lo mismo vale para las omisiones: Si alguien tiene derecho de abstenerse de hacer tal cosa, es ilícito forzarlo a hacer esa cosa. En definitiva, lo que caracteriza un derecho para alguien es la obligación (derivada) para todos los demás de no impedir ni obstaculizar el ejercicio de tal acción. Hemos llamado a este postulado, principio de obligación sobrevenida: Si haciendo p uno impide a alguien hacer q (lo fuerza a no q), entonces en la medida en que ese alguien tiene derecho a hacer q , está prohibido p (es decir, es un deber no hacer p) [$p \text{ f } Nq \ \& \ Pq \rightarrow ONp$]. Lo que caracteriza al titular de un derecho es que los demás están obligados a no estorbar su ejercicio.

También fundamental para un adecuado tratamiento de los derechos positivos es tomar en consideración el papel de los cuantificadores, especialmente del cuantificador existencial. Ya hemos dicho que una caracterización de los derechos positivos es aquella que los define como los derechos en cuya formulación lo que viene después del operador deóntico “es lícito que” se inicia con un cuantificador existencial. Sin embargo, tradicionalmente y salvo contadas excepciones, la lógica deóntica estándar ha dejado de lado las consideraciones sobre la cuantificación en los contextos

⁷ Este principio puede rastrearse en la literatura jusfilosófica desde los juristas romanos y medievales hasta autores contemporáneos como Bobbio (1956[1991], p. 97):

“... una norma que impone un deber a una persona atribuye, a mismo tiempo, a otra persona el derecho de exigir el cumplimiento, así como una norma que atribuye un derecho impone, al mismo tiempo, a otros el deber de respetar el libre ejercicio o de permitir la ejecución.”

normativos⁸. A nuestro entender, un adecuado tratamiento lógico de las relaciones entre normas debe incluir cuantificadores, como se pone de manifiesto en el caso del análisis de los ‘derechos a un algo’: al trabajo, a atención médica, a una vivienda, a educación, etc. Para ello tomamos como básico el siguiente ‘principio de determinación deontica’: El único modo de hacer obligatorio que algo sea así o asá es establecer para alguna entidad particular el deber de que sea así $[O\exists xp \rightarrow \exists xOp]$. Ningún sistema normativo puede imponer la obligación de que al menos una entidad sea r sin considerar -implícita o explícitamente- que existe al menos una entidad bajo tal deber r . Alternativamente, en la medida en que toda entidad es tal que puede lícitamente tener cierta característica, entonces es lícito para todas las entidades tomar tal característica $[\forall xPp \rightarrow P\forall xp]$. Hay que señalar que los correlatos modales de estos postulados no son válidos, mientras que los equivalentes deonticos de los axiomas modales clásicos referidos a la cuantificación resultan equivocados desde un punto de vista del razonamiento normativo, abundando así en la profunda discrepancia entre la ‘posibilidad deontica’ y otros tipos de posibilidad⁹.

Algunos otros principios de nuestra lógica deontica son el de obligación disyuntiva (una obligación disyuntiva implica que, al menos, uno de los disyuntos es obligatorio); el principio llamado de Kant (lo que es inevitablemente obligatorio es entonces factible o realizable; y los principios del silogismo disyuntivo deontico (cuando una situación p no sucede en absoluto, entonces la licitud (respectivamente la obligatoriedad) de una disyunción entre p y q implica la licitud (respectivamente la obligatoriedad) de q).

Lógica deontica y derechos positivos

La lógica deontica va a resultar fundamental para permitirnos establecer el alcance de los cuantificadores existenciales involucrados en la formulación de los derechos positivos. Partiendo de los principios deonticos antes formulados, se obtienen algunos resultados clarificadores en relación al análisis de los derechos positivos. En los siguientes corolarios subyace la interpretación del cuantificador existencial como una suerte de disyunción generalizada, una disyunción de enunciados singulares que tendrá tantos elementos como elementos tenga el conjunto que constituye el campo de valores de la variable $[\exists xr \leftrightarrow r_1 \vee r_2 \vee \dots \vee r_n]$.

⁸ Hintikka será el precursor del uso de cuantificadores para la lógica deontica (Hintikka 1957). Pero serán muy pocos los análisis lógicos de las calificaciones deonticas que incluyan elementos cuantificacionales. Entre los más recientes, véase Goble (1996).

⁹ Ausín y Peña (1995, pp. 227-228).

1º En la medida en que alguien tiene un derecho positivo a algo (p.e. a una vivienda), entonces existe al menos una cosa de ese tipo tal que está prohibido impedirle acceder a ella.

Sean 'p': Luis accede a una vivienda x , y 'q': Luis no es obstaculizado para acceder a una vivienda x . La relación de forzar, impulsar, constreñir la representamos como f , y la negación como N .

(2) $P\exists x p$	Luis tiene el derecho positivo a una vivienda
(3) $N\exists x q \ f \ N\exists x p$	asunción empírica
(4) $PN\exists x q \rightarrow ON\exists x p$	(3), principio de obligación sobrevenida
(5) $NON\exists x p$	(2), df.P
(6) $NP N\exists x q$	(4), (5), cálculo cuantificacional
(7) $O\exists x q$	(6), df.O
(8) $\exists x Oq$	(7), principio de determinación deóntica

2º Si uno no puede acceder a ninguna otra vivienda, entonces, en la medida en que tiene derecho a que haya una vivienda u otra a la que tenga acceso, tiene un legítima reclamación sobre esta vivienda particular¹⁰. Es lo que hemos denominado, principio de legítimas demandas.

Por tanto, lo que se demuestra es que en los sistemas legales que formulan el derecho positivo para todo el mundo a tener una vivienda (o educación, o asistencia médica, etc.), se infiere una prescripción que garantiza el beneficio personal no accesorio sobre cada vivienda particular, siempre que se den dos condiciones: (1) que el demandante sin techo no pueda acceder a cualquier otra vivienda; y (2) que no haya ningún otro alojamiento disponible. Así pues, quien teniendo el derecho positivo a una vivienda y careciendo de tal, ve una vivienda desocupada, puede reclamar legítimamente esa vivienda. El derecho a una vivienda comporta el deber de los demás de no estorbar el acceso de cada uno a una vivienda. Eso acarrea que haya alguna vivienda tal que la autoridad tiene la obligación de no estorbar el acceso a esa vivienda de quien lo necesita, por lo que la 'ocupación' nos parece claramente justificada, especialmente cuando se refiere a edificios vacíos a menudo destinados a operaciones especulativas.

Si no se admitiera ninguna de estas implicaciones el tal derecho positivo sería papel mojado, mera retórica de adorno en los textos legales. Precisamente la lógica deóntica nos proporciona una sólida respuesta acerca de los compromisos y exigencias que comporta el reconocimiento de derechos positivos.

Pero esto supone que, cuando los sistemas legales garantizan nominalmente derechos positivos pero no ponen los medios materiales necesarios para asegurar tales derechos, se producen graves conflictos: Por un

¹⁰ Los detalles de esta demostración pueden consultarse en Ausín y Peña (en prensa).

lado, un individuo o una colectividad pueden poseer un bien X y, consiguientemente, tener derecho a impedir a los demás que accedan a ello; por otro lado, puede suceder -y así sucede a menudo- que a la gente D se le otorga el derecho cuantificacional a un bien del tipo de X pero no puede acceder a ningún otro de ese tipo. Entonces, en virtud del principio de no impedimento, no se debe impedir su acceso a X en concreto; o si se prefiere, se tiene la obligación sobrevenida de no obstaculizar el acceso a X por D. (Por descontado que hay muchos casos intermedios).

Tomando los derechos positivos en serio

Parafraseando el afamado libro de Dworkin (*Taking Rights Seriously*) vamos a establecer a continuación algunas conclusiones sobre el carácter de los derechos positivos, su reconocimiento y las consecuencias que conllevan.

1.- Los legisladores son responsables de articular, en la medida de lo posible, un orden social donde los conflictos anteriores no sucedan. En consecuencia, la gente pobre y sin recursos tiene el derecho a demandar subsidios sociales, inversiones, redistribución del uso de viviendas, etc. para atajar al menos los problemas sociales más acuciantes. Reclamación que ha de darse tanto ante el Estado como ante los particulares, en la medida en que va cuajando la idea de que los derechos fundamentales son derechos *erga omnes*. Y si la comunidad internacional (ONU, GATT, etc.) no fuerza a los Estados miembros a adoptar políticas en este sentido, esa comunidad está actuando ilegalmente, pisoteando el propio derecho internacional que ella misma ha aprobado. Si bien puede considerarse que la Declaración univesal de derechos humanos de 1948 no pasa de ser una normativa moral de carácter muy amplio, sí se considera que los Pactos que la desarrollan son de carácter vinculante para cada uno de los Estados que los ratifiquen (tal es el caso del Pacto sobre los derechos económicos, sociales y culturales de 1966)¹¹.

2.- El respeto y cumplimiento de los derechos positivos conlleva una serie de obligaciones sobrevenidas que pueden entrar en contradicción con otros derechos (negativos) individuales, algunos de ellos también reputados como fundamentales. En lo dicho hasta ahora, esto es evidente para el derecho de propiedad. Los derechos a la atención médica, a la vivienda, a la seguridad social, a un medio ambiente no contaminado, etc. acarrear -por el principio de no interferencia- la prohibición para otros de poner estorbos que impidan u obstaculicen a los demás la adquisición de estos bienes. Por tanto, a nuestro entender, el derecho de propiedad privada, si existe, habrá de ser restringido y estar subordinado a que lo privadamente poseído se administre de manera que se haga posible el disfrute de bienes básicos por los demás (atención sanitaria, alimentación, educación, vivienda, etc.). En la medida en

¹¹ En 1976 se alcanzó el número necesario de ratificaciones (35 adhesiones) para que el Pacto adquiriese valor jurídico, entrando en vigor el 3 de enero de dicho año.

que la posesión particular no propicie ese logro, o lo obstaculice, o incluso lo impida, el derecho de propiedad habrá de ceder paso a los deberes dimanantes de los derechos positivos de cada uno. En definitiva, tratase de una cuestión de prioridad entre ciertos derechos negativos y los derechos positivos. Si hay compatibilidad, mejor, pero en el conflicto entre el derecho a la propiedad privada y el derecho a la satisfacción de las necesidades socio-económicas y culturales, nos inclinamos por éstos últimos derechos positivos. Ya admitía el “revolucionario” Tomás de Aquino que en caso de necesidad, se vuelven caducos o nulos los derechos de propiedad que obstaculicen la satisfacción de tales necesidades.

3.- El fundamento de esta reflexión presupone una toma en serio de los derechos positivos, especialmente de aquellos del ámbito socio-económico y cultural, solemnemente reconocidos en el derecho internacional. Sin embargo, esto no sucede en España, donde hay una devaluación jurídica de los derechos sociales, merced a que el art. 53 de la Constitución de 1978 (C.E.) los convierte, como mucho, en un principio que “informará la legislación positiva, la práctica judicial y la actuación de los poderes públicos”. Este epígrafe se refiere en concreto a los “derechos” reconocidos en el Capítulo tercero de la misma: art. 39: protección social, económica y jurídica de la familia; art. 40: progreso social y económico y distribución de la renta regional y personal más equitativa; art. 41: seguridad social (asistencia y prestaciones); art. 43: salud; art. 44: cultura; art. 45: medio ambiente; art. 47: vivienda digna y adecuada, impidiendo la especulación; art. 50: pensiones¹². La consecuencia es que no cabe recurso de amparo ante el Tribunal constitucional en estos casos; estos “derechos” sólo podrán ser alegados ante la Jurisdicción ordinaria de acuerdo con lo que dispongan las leyes que los desarrollen. Así, aunque superficialmente estos preceptos están encuadrados dentro del Título I de la C.E. (De los derechos y deberes fundamentales) son definidos sin arrobo alguno como “no derechos fundamentales”¹³. De este modo, al “promulgar” derechos positivos vacíos de contenido se burla el derecho internacional y se perpetra un engaño hacia los ciudadanos.

4.- Se nos plantea en este punto la pregunta sobre si debe ser castigada la ausencia de acciones para satisfacer un derecho positivo (como así sucede en el caso de la denegación de auxilio). Por un lado, la lógica deóntica nos

¹² No obstante pueden plantearse situaciones conflictivas: El art. 19 reconoce el derecho de los españoles a elegir libremente su residencia y está enmarcado entre los derechos fundamentales *fetén*. Sin embargo, el que no se reconozca como fundamental el derecho positivo a una vivienda conlleva que no se garantice la mencionada libertad de residencia, ya que para obtener la vencidad administrativa en cualquier sitio de España es preciso un domicilio, del tipo que sea. Pero ¿qué sucede si no se tiene ningún domicilio, si no se puede acceder a ningún alojamiento?

¹³ Igualmente denominados como “derechos de configuración legal pero no constitucional” (Pérez Royo, 1994, p. 242).

dice que, si el no poner los medios para satisfacer un derecho positivo impide su realización, esa omisión está prohibida y, subsiguientemente, es un deber la acción correspondiente. En ese sentido, el incumplimiento de una obligación podría considerarse como una falta susceptible de penalización. Sin embargo, un principio básico del orden jurídico, el principio de legalidad, nos dice que una conducta no constituye un delito a menos que previamente haya sido así considerada por la ley (*nullum crime sine lege*). A nuestro entender, una interpretación extensiva de ese delito de denegación de auxilio cabría para la omisión de auxilio en los casos de las tremendas hambrunas, catástrofes, problemas sanitarios, subdesarrollo en definitiva -sobre todo hacia aquellos países e individuos con mayor desarrollo económico. Aunque reconocemos que el encaje explícito de este delito en el derecho internacional puede ser complicado, su fundamento está plenamente justificado atendiendo a la lógica deóntica: Si los Pactos internacionales reconocen una serie de derechos positivos que conllevan un conjunto de deberes, de obligaciones derivadas para la comunidad internacional, entonces su incumplimiento debería ser castigado y reparado por esa misma comunidad. Constituye un tremendo agravio hacia la mayor parte de la humanidad ese incumplimiento de los derechos económicos, sociales y culturales, indispensables a su dignidad y al libre desarrollo de su personalidad (art. 22 de la Declaración universal de derechos humanos).

“... la prosperidad económica no consiste tanto en la abundancia total de bienes, sino más bien en la *real y eficaz distribución* según la justicia, de forma que todos los ciudadanos puedan perfeccionarse, en lo que consiste el verdadero fin de la economía.” Juan XXIII, *Mater et Magistra*.

BIBLIOGRAFÍA

- Alchourrón, C.E. y Bulygin, E.: “Permisos y normas permisivas”, en *Análisis lógico y derecho*. Madrid: Centro de Estudios Constitucionales, 1991: 215-238.
- [Díez] Ausín, F.J. y Peña, L.: “Un sistema de lógica deóntica sin principio de cierre”, en *Actas del I Congreso de la SLMFC*. Madrid: UNED, 1993: 44-47.
- Ausín, F.J. y Peña, L.: “Quantificational Entitlements and Relevantoid Deontic Logic”, *Logique et Analyse* 150-151-152, 1995: 209-238.
- Ausín, F.J. y Peña, L.: “Paraconsistent Deontic Logic with Enforceable Rights” (en prensa).
- Bobbio, N.: *Teoría general del derecho*. Madrid: Debate, 1991.
- Contreras, F.J.: “Derechos sociales y necesidades básicas: Respuesta al profesor De Castro”, *Anuario de Filosofía del Derecho* XII, 1995: 689-704.
- De Castro, B.: “Caracterización y fundamentación de los derechos sociales. Reflexiones sobre un libro de F.J. Contreras”, *Anuario de Filosofía del Derecho* XII, 1995: 679-687.
- Goble, L.: “Ought and Extensionality”, *Noûs* 30(3), 1996: 330-355.
- Hintikka, J.: “Quantifiers in Deontic Logic”, *Societas Scientiarum Fennica* XXIII.4, 1957: 3-23.
- Peña, L.: “A Chain of Fuzzy Strengthenings of Entailment Logic”, en *Actas del III Congreso español de tecnologías y lógica fuzzy*. Santiago de Compostela: Universidad de Santiago, 1993: 115-122.

Peña, L.: "El bien común, principio básico de la ley natural", *Isegoria* 17, 1997: 137-163.

Pérez Royo, J.: *Curso de derecho constitucional*. Madrid: Marcial Pons, 1994.

Sánchez-Mazas, M.: *Cálculo de las normas*. Barcelona: Ariel, 1973.

Lorenzo Peña

Instituto de Filosofía. CSIC

e-mail: lorenzo@ifs.csic.es

F. José Ausín

CALIJ

e-mail: txetxu@sf.ehu.es



RATIONAL DELIBERATION AND REACTIVE ATTITUDES

MONIKA BETZLER

Georg-August University Göttingen

ABSTRACT: After a well reasoned choice between two conflicting options we often display emotional attitudes like regret. Rational deliberation is alleged for not being able to take account of our emotional phenomenology. If we really had such good reasons for our choice, we didn't have the need to regret any losses. An increasingly widespread view tries to accommodate the fact of reactive attitudes without giving up on the normative force of rational deliberation. It is maintained that values are incommensurable, and therefore reactive attitudes are asked for. I will first discuss the claim of value incommensurability and its explanatory force with respect to our emotions. I will then show that this attempt fails. Finally, I present my own view on the rationality of regret by analyzing its propositional content. If we didn't regret we couldn't have a practical identity.

KEYWORDS: Rational choice, practical deliberation, reactive attitudes, emotions, regret, value incommensurability, practical identity.

1. The Problem of Rational Choice between Conflicting Options

Our life is such that we often face choices between conflicting options. They prove incompatible because they cannot both be chosen and acted upon. However, they represent ends we both value. We would be ready to choose each of them if they didn't coincide. When we are confronted with a choice between two options A and B that promote different ends, we choose – given that we cannot perform both under the circumstances – the option which we consider to be supported by weightier and thus better reasons. At least so it seems.

However, often times it proves to be rather difficult to make out these “better reasons”. Particularly in cases, where the two options between which we have to choose, both either have strong moral implications (like “moral dilemmas”) or significant, i.e. far-reaching consequences for our life (like a choice between different careers) we can provide more often than not good

reasons for both of them without being able to assess which ones are actually weightier, or whether there are weightier reasons at all. Moreover, our psychology indicates that there is something worrisome about such choices: we tend to display emotional attitudes *after* such a choice as well deliberated as it might have been since we had to forgo another equally (or at least highly) valued option.¹ Even if weightier reasons support option A over option B, and hence, even if we seem to value option A more than option B, we *regret* that we had to renounce option B.

Let's think of some examples. In the case of a moral dilemma we face two conflicting obligations.² Imagine the following scenario:

(1) A friend of mine promised me to comment on my paper I plan to deliver in Maribor, and help me with the last revisions. On her way to my house, however, she witnesses a car accident. She breaks her promise in order to rescue the victim. Her actual choice seems to be supported by weightier reasons so that her decision to break the promise is well justified. But she still might have a bad conscience and regret to have made me wait in vain for her to come and help, particularly if she thinks how much better my paper could have been with her assistance.

(2) Here's a tragic case: a parent is only able to rescue one of her two children out of a burning house. She has to select one child, and forgo her obligation toward her second child. We would be appalled if the parent didn't feel deep distress after that choice despite the fact that it was well justified. There simply was no possibility to rescue both of the children in the given circumstances.

In the case of a choice between options that have a significant impact on our life we can also face conflicts.³

(3) Think of the question we all had to confront at a certain age, namely, what career we want to pursue. Given the uncertainties about how various professions really turn out to be, and in light of the difficulty how to assess what we really value, such a choice can be traumatizing. But even if we opt for a career path that seems justified by virtue of the information available at the time of the choice – let's say we opt for medicine –, we might have to put up with regret. We might regret that we had to reject another highly valued option, for example, the career as a philosopher. There is also the

¹ B. Williams was one of the first to point to the importance of regret. Cf. B. Williams (1965), 115-137; (1980), 20-39; (1985), 177.

² Cf. C. Gowans (1987); C. Gowans (1994); W. Sinnott-Armstrong (1988); P. Greenspan (1995); H. Mason (1996).

³ Cf. J. Raz (1986); R. Chang (1997).

possibility that we will deeply regret our choice if the career as a physician, for unforeseeable reasons, does not meet our expectations.

These examples obviously differ in various aspects. But in all three of them we regret losses that occurred as a consequence of a seemingly well justified choice. Rational deliberation is therefore alleged for not being able to take account of our emotional phenomenology. If we really had such good reasons for our choice, we didn't have the need to regret any losses. Therefore, the question arises whether our choices in light of conflicting options can be well justified at all. By expressing reactive attitudes we seem to question normative concepts of practical reasoning and their action guiding function.⁴ They are unable to take into account that the options not acted upon still represent values we cherish, or obligations we feel committed to, and therefore, at least in some way, seem to remain in force.

According to a utilitarian account, for example, it would be rational to choose the option that had the higher utility function, and forget about the option with the lower utility function.⁵ Equally, it would be rational by Kantian lights to choose the option which wins the universalizability test. If two obligations seem to conflict, only one turns out to be our "all things considered" obligation. The other one simply is not an obligation at all.⁶ It is not the one that we can will to be a universal law given the circumstances. Even in tragic cases it seems rational to choose one option – since we can't do both – and simply accept that the other option has to be neglected.

Given our reactive attitudes, however, normative theories miss out on something.

But what do reactive attitudes like regret or a deep sense of loss exactly imply for practical reasoning? What do we actually react to in displaying such emotional and painful attitudes?

An increasingly widespread view tries to accommodate the fact of reactive attitudes without giving up on the normative force of rational deliberation. It is maintained that *values are incommensurable*, and therefore reactive attitudes are asked for.

In the following, I will *first* discuss the claim of *value incommensurability* and its explanatory force with respect to our emotions. I will *then* show that this attempt to make sense of reactive attitudes fails. *Finally*, I will venture to present my own view on their rationality by analyzing their propositional content in more detail.

⁴ Cf. J. Raz (1986), 334.

⁵ Cf. R. Hare (1981), 206n.

⁶ Cf. A. Donagan (1984), 271-290; (1993), 11-22; B. Herman (1993), 167n; T. Hill (1996), 167-198.

2. Incommensurability of Values

The claim of value incommensurability is supposed to explain our reactive attitudes.⁷ If ends that we both value are taken to be incommensurable, we seem to be justified in taking evaluative attitudes to each of the ends even if we can only choose one. But what exactly is value incommensurability?

The options and their promoted valuable ends among which we have to choose are taken to be incommensurable, if *neither of them is better than the other* with respect to a common value. Hence, valuable ends cannot be compared, because they cannot be measured on a scale.⁸ Moreover, incommensurable values are *not of equal value*. They are at best *roughly equal*. Rough equality implies that we are indifferent between two options, since they are supported by weighty but totally different, and therefore incomparable reasons. The options are only equal with respect to their significance to us. However, they are significant for very different reasons.⁹ Therefore, there could be a third option, that is better than the first but not better than the second option. This is supposed to show that the first two options are not equal in value.¹⁰

As Ruth Chang puts it: "If items are incomparable, nothing affirmative can be said about what value relation holds between them."¹¹

Given that values are incomparable *rational choice* can take on the following meaning: reason can determine how we should act, but we can justify our choice only for *undefeated reasons*, but not for reasons "all things considered". Consequently, it seems possible that we divide our attitudes *in proportion* to the options' relative, but incomparable values: one is foregone and therefore regretted, the other one is promoted and therefore enjoyed.¹²

There remain valuable options unacted upon that are supported by weighty, and very different reasons. The value of an option is taken to be independent of whether or not it can be chosen. The reason because of which we finally opt for one action does not necessarily defeat all the others. Instead,

⁷ I have to leave other claims like dilemma realism as well as attempts of utilitarianists and Kantians to amend their theories for reasons of space aside.

⁸ In that sense incomparability is even the stronger claim than incommensurability. According to incommensurability in the strong sense of the word two options cannot be measured on a single scale. I will use the term 'incommensurability' in the following as incomparability.

⁹ Cf. J. Raz (1986), 331n; J. Griffin (1986), 96n.

¹⁰ Cf. J. Raz (1986), 324.

¹¹ R. Chang (1997), 4.

¹² Cf. T. Hurka (1996), 558. Hurka, however, thinks that a modest pluralism can account for regret.

the agent is confronted with a range of eligible options, and it is rational to choose either of them for an undefeated reason, but without comparison.¹³

In the case of moral dilemmas conflicting obligations are traced back to incommensurable values like obligations, rights, utility, perfectionist goals and personal commitments that supposedly cannot be compared with each other.¹⁴ In the case of significant choices between comprehensive goals the conflicting options represent incomparable values like – if you think of a choice between medicine and philosophy – between devotion to others or financial security versus, let's say, intellectual fulfillment or simple leisure.

The underlying assumption of this claim is that we could not make sense of our reactive attitudes if the values promoted by the conflicting options were comparable. There would be no reason, so it is argued, to regret an option and its forgone value if this value could be compared with the value for the sake of which we came to choose the other option. If the foregone value, however, cannot be compared with the value we opted for, we hold different reasons for different, i.e. incommensurable values. We can hold reasons that support various incomparable options. The reasons for the option not acted upon still prevail, thus providing ground for reactive attitudes.

However, while the view that values are incommensurable might take account of regret, it remains pretty modest about the role of practical deliberation. If there is more to practical deliberation and values therefore prove to be comparable, reactive attitudes remain unexplained.

3. Value Incommensurability Doomed: The Task of Practical Deliberation

I have to content myself with the outline of three problems which the claim of value incommensurability entails. Admittedly, my critique will remain sketchy, but I hope to highlight some concerns that this claim gives rise to.

First of all, it rests on a wrong assumption: comparability is taken to imply that the forgone value is *included* in the value opted for ("inclusion view"). According to this interpretation nothing would be really forgone at all. Consequently, there would be no reason left to regret anything. But this seems false to the facts.

The value of option A could only be included in the value of option B if both can be compared on a single scale (e.g. utility, pleasure, security, etc.)

¹³ Cf. J. Raz (1986), 338n.

¹⁴ Cf. T. Nagel (1979), 128-141; Lemmon (1962), 139-158; B.C.v. Fraassen (1973), 138n; P. Railton (1992), 736.

Comparability of values, however, does not necessarily imply measurement on a single scale. We can compare different values in various ways, and on different scales.

Second, it assigns a wrong cause to the fact that some choices are difficult to make. The difficulty is not due to the fact that the options between which we have to choose are incomparable. Think of the choice between two careers. It is neither the life of a physician as such that is incomparable with the life of a philosopher. Nor is it the fact that – in the case of moral dilemmas – some of our obligations are due to personal relations, others due to considerations of, let's say, utility. Instead, the difficulty is to find out what we *really value more*, or what we are *really obligated* to do in the case at hand.

Third, it misunderstands the role of practical deliberation before a choice takes place. It is the specific task of practical deliberation to *describe* the situation of choice such that we can perceive the various reasons that speak for or against the options in question.¹⁵ In light of the dispositions, the specific biography of a certain person, and other available information about the respective goals, we are usually able to decide which one is better justified “all things considered”. Practical deliberation guides our actions by providing us with a *decision frame* that helps us characterize the choice at hand such that relevant aspects of various options can be perceived and compared. We do not content ourselves with the mere description of various options (- and that's what the claim of value incommensurability does -), but try to relate them in their significance to the given situation, the persons affected, the possible alternatives etc.

Even if there is no single scale which allows to commensurate values, there are various aspects of comparability which we consider in a choice, and which we feed into our deliberation before a choice “all things considered” takes place. Practical deliberation, however, is focused on guiding our actions.¹⁶ It cannot accommodate the fact of reactive attitudes, since they arise if a highly valued option is unacted upon. Therefore, we must look for another explanation of our reactive attitudes. The claim of value incommensurability, in any case, seems mistaken.

¹⁵ E. Ullmann-Margalit and S. Morgenbesser (1977), 775n emphasized that choices are dependent on how we describe them. Cf. E. Anderson (1993), 23n.

¹⁶ M. Stocker (1990), 110n, emphasized the importance of non-action-guiding act evaluations that normative theories cannot handle.

4. The Rationality of Reactive Attitudes: The Quest for Practical Identity

We can make sense of reactive regret by analyzing its cognitive structure, i.e. its propositional object. I maintain that regret is appropriate even if the claim of value incommensurability is unable to account for it.

To be susceptible to regret in one's relation to an executed choice is connected to a distinctive form of evaluative belief. It is the belief expressing the propositional object of regret, i.e. that a moral or non-moral expectation one holds oneself to has been violated. Contrary to remorse or guilt, regret is not accompanied by the belief that we acted wrongly. Instead, we respond to the fact that for contingent reasons we were not able to stick to our ideals and norms.¹⁷

In the moral domain, an expectation one holds oneself to is a distinctively moral end like an obligation. In the non-moral domain one holds oneself to values, ideals or norms one accepts as guiding one's life. If we can compensate the forgone values, the reactive attitudes will eventually disappear. E.g., if we excuse ourselves for having had to break a promise in order to help an accident victim, and the person accepts our excuse, we will stop regretting. Our excuse serves to show that we had good reasons to break an obligation we generally commit ourselves to. However, if the values not acted upon cannot be compensated – e.g. in tragic cases or cases with great significance for our entire life – our reactive attitudes might attenuate over time, but they will remain. But what is it that makes them rational or appropriate?

If we did not confront us with the expectations we hold ourselves to, we could neither live a coherent life nor be motivated to do so. And this shows something about our moral life too. We do not only rationally deliberate from a first person perspective about our *future* actions by comparing various ends as difficult as that comparison might already be. Once we have made choices – and usually we cannot evade them without major losses – we evaluate in retrospect our past choices with respect to their significance for ourselves as well as for others. Even if we were not responsible for the outcomes of our choices – often times we just didn't have a choice to pursue all the ends we valued – we have to take into account the meaning of these losses for us and others. This doesn't mean that we are guilty for not having chosen the goals we simply were not able to choose.¹⁸ But we have to make sense of these losses in light of our own norms and ideals that were breached. We can even say: if we didn't display reactive attitudes like regret we couldn't have a

¹⁷ Cf. J. Landman (1993), 45; R.J. Wallace (1994), 19n.

¹⁸ C. Gowans (1994) defends this view. Cf. P. Greenspan (1995).

practical identity.¹⁹ We regard our choices as expressive of ourselves. The conception of our practical identity is a description under which we find our life worth living and our choices worth undertaking. In that sense the reflective structure of human consciousness establishes a relation which we have to ourselves. In order to be coherent with our practical identity that gives rise to reasons and obligations we have to take a stance to events that do not cohere with that conception.

My conclusion is simple: in practical deliberation we compare various valued options with respect to aspects that are relevant for the choice at hand. In reactive attitudes, however, we take account for the meaning the forgone values have for our conception of practical identity.

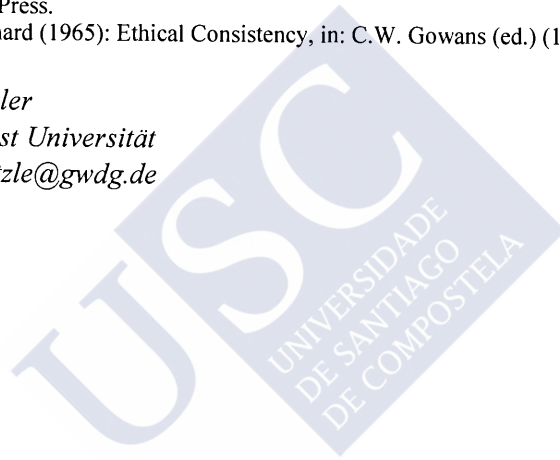
REFERENCES

- Anderson, Elizabeth (1993): *Value in Ethics and Economics*. Cambridge/Mass. and London: Harvard University Press.
- Chang, Ruth (1997): Introduction, in: (1997), 1-34.
- Chang, Ruth (ed.) (1997): *Incommensurability, Incomparability, and Practical Reason*. Cambridge/Mass. and London: Harvard University Press.
- Donagan, Alan (1984): Consistency in Rationalist Moral Systems, in: C.W. Gowans (ed.) (1987), 271-290.
- Donagan, Alan (1993): Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy, in: H.E. Mason (ed.) (1996), 11-22.
- Fraassen, Bas C.v. (1973): Values and the Heart's Command, in: C.W. Gowans (ed.) (1987), 138-153.
- Gowans, Christopher W. (1994): *Innocence Lost*. An Examination of Inescapable Moral Wrongdoing. Oxford and New York: Oxford University Press.
- Gowans, Christopher W. (ed.) (1987): *Moral Dilemmas*. New York and Oxford: Oxford University Press.
- Greenspan, Patricia S. (1995): *Practical Guilt*. Moral Dilemmas, Emotions, and Social Norms. New York and Oxford: Oxford University Press.
- Griffin, James (1986): *Well-Being: Its Meaning and Measurement*. Oxford: Oxford University Press.
- Hare, Richard M. (1981): *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon.
- Herman, Barbara (1993): *The Practice of Moral Judgment*. Cambridge/Mass. and London: Harvard University Press.
- Hill, Thomas E. (1996): Moral Dilemmas, Gaps, and Residues: A Kantian Perspective, in: H.E. Mason (ed.) (1996), 167-198.
- Hurka, Thomas (1996): Monism, Pluralism, and Rational Regret, in: *Ethics*, 555-575.
- Korsgaard, Christine (1996): *The Sources of Normativity*. Cambridge et al.: Cambridge University Press.
- Landman, Janet (1993): *Regret*. The Persistence of the Possible. New York and Oxford: Oxford University Press.
- Mason, Homer E. (ed.) (1996): *Moral Dilemmas and Moral Theory*. New York and Oxford: Oxford University Press.
- McConnell, Terrence C. (1996): Moral Residue and Dilemmas, in: H.E. Mason (ed.) (1996), 26-47.
- Nagel, Thomas (1979): The Fragmentation of Value, in: C.W. Gowans (ed.) (1987), 174-187.

¹⁹ Cf. C. Korsgaard (1996), 101. I thank C. Korsgaard for discussion on this issue.

- Railton, Peter (1992): Pluralism, Determinacy, and Dilemma, in: *Ethics* 102, 720-742.
- Raz, Joseph (1986): *The Morality of Freedom*. Oxford: Clarendon.
- Raz, Joseph (1997): Incommensurability and Agency, in: R. Chang (ed.), 110-128.
- Scanlon, Thomas (1986): The Significance of Choice. *The Tanner Lectures of Human Values*, 151-216.
- Sinnott-Armstrong, W. (1988): *Moral Dilemmas*. Oxford: Basil Blackwell.
- Stocker, Michael (1990): *Plural and Conflicting Values*. Oxford: Clarendon.
- Ullmann-Margalit, Edna and Sidney Morgenbesser (1977): Picking and Choosing, in: *Social Research* 44, 757-785.
- Wallace, R. Jay (1994): *Responsibility and the Moral Sentiments*. Cambridge, Mass. and London: Harvard University Press.
- Williams, Bernard (1980): Moral Luck, in: *Moral Luck*. Cambridge: Cambridge University Press 1981, 20-39.
- Williams, Bernard (1985): *Ethics and the Limits of Philosophy*. Cambridge: Harvard University Press.
- Williams, Bernard (1965): Ethical Consistency, in: C.W. Gowans (ed.) (1987), 115-137.

Monika Betzler
Georg-August Universität
e-mail: mbetzle@gwdg.de





A PLEA FOR A NEW ANALYTICAL INTERPRETATION OF G.E. MOORE'S ETHICAL PHILOSOPHY

MARIA MICHELA MARZANO-PARISOLI
Centre de recherche Sens, éthique et société - CNRS

ABSTRACT: The aim of this paper is to point out that the traditional analytical interpretation of Moore's ethics is not able to give an account of the metaphysical backdrop of Moore's position. In order to better understand the facts/values relation in Moore's works, this paper analyses the intrinsic value/intrinsic nature relation and the intrinsic nature/intrinsic properties relation as ontological relationships. Intrinsic properties are in fact substantial and essential properties and give to the object all the substance that it has. Intrinsic value is not an essential property: it depends on the intrinsic nature of a thing, but it can never be reduced to the intrinsic properties of this same object.

KEYWORDS: Essentialism, dependent property, good, independent property, intrinsic nature, intrinsic property, intrinsic value, reductionism.

1. Introduction

After the publication of *Principia Ethica* in 1903 (PE)¹, moral analytical philosophy has always tried to clarify the question of moral terms' meaning as well as the problem of the naturalistic fallacy. All at once it seemed as if the most metaphysically attractive theories of morality were not able to solve the most important problems of morality because of their incapacity to correctly analyse the meanings of moral terms and to avoid the naturalistic fallacy. Nevertheless, the interpretation of the naturalistic fallacy has often been a too narrow and poor one so that now we probably could not be able to give a correct account of Moore's position if we continue to think at the naturalistic fallacy as the fallacy of trying to define good, which is an unanalysable moral term. As in fact Moore himself writes in his *Preface* to the second edition of PE (1922): "whether G is analysable or not, what is more important is to insist that this question has nothing like the importance which I attributed to it"². According to Moore, the real meaning of the naturalistic fallacy is that G

¹ Moore G.E., *Principia Ethica*, Cambridge University Press, Cambridge, 1st ed. 1903. References are to revised ed., edited and with an introduction by Thomas Baldwin (Cambridge University Press, Cambridge, 1993).

² Moore G.E., "Preface to the Second Edition", PE, pp. 2-27. It was written in 1922, but was left unpublished for over seventy years.

(i.e. good) is a kind of intrinsic value which we can never reduced to any intrinsic property.

The aim of this paper is to point out that the traditional analytical interpretation of Moore's ethics is not able to take an account of the subtle but also strong ontology of Moore's position³. If we really want to understand the correct relation between facts and values in Moore's sentiment, we have to analyse either the relationship between G and the intrinsic nature of a thing which possesses it, or the intrinsic nature/intrinsic properties relation. In this sense, my hypothesis is that we really can arrive at understanding the ontological meaning of Moore's fact/value gap only if we correctly read these two relations. This is why I would like to propose an ontological interpretation of Moore's ethical position by means of an analysis either of the most important differences between intrinsic properties and intrinsic value or of the relationship between intrinsic properties, intrinsic value and intrinsic nature of a thing. But in order to explain this, I have to refer not only to PE, but also to some other Moore's works such as the revised *Preface*, and *The Conception of Intrinsic Value* (CIV)⁴.

II. The natural/non natural distinction among properties

It is not very simple to understand what are according to Moore natural and non-natural properties. As Moore himself recognises in his revised *Preface*: "my attempts to define natural properties are hopelessly confused"⁵. Nevertheless, in order to understand Moore's position, we have to examine the § 26 of PE, where Moore discusses the problem of which among the properties of natural objects are natural properties and which are not. As Baldwin has clearly pointed out, Moore's answer is a strange one. For this reason, I will give Moore's answer in full and then I will discuss these passages also by means of Baldwin's interpretation⁶: "My test for these [i.e. natural properties] too also concerns their existence in time. Can we imagine 'good' as existing *by itself* in time, and not merely as a property of some natural object? For myself, I cannot imagine it, whereas with the greater number of properties of objects - those which I call the natural properties - their existence does seem to me to be independent of the existence of those objects. They are rather parts of which the object is made up than mere predicates which attach to it. If they were all taken away no object would be

³ Cf. O' Connor D., *The Metaphysics of G. E. Moore*, Reidel Publishing Company, Dordrecht, Boston, London, 1982.

⁴ "The Conception of Intrinsic Value", in *Philosophical Studies* (Routledge, London, 1992²). References are here to the paper reprinted in the revised edition of PE.

⁵ Moore G.E., "Preface to the Second Edition", PE, p. 13.

⁶ Baldwin T., "Ethical Non-Naturalism", in Hacking I. (ed.) *Exercises in Analysis*, Cambridge University Press, Cambridge, 1985, pp. 24-45.

left, or even a bare substance: for they are in themselves substantial and give to the object all the substance that it has. But this is not so with good"⁷.

But how could we understand these passages? First of all we have to distinguish three different theses, that is to say three different explications that Moore gives us about a natural property. *Thesis I*: a natural property is a property whose existence does seem to be independent of the existence of those objects which possess it. *Thesis II*: a natural property is a part of which the object is made up rather than a mere predicate which attaches to it. *Thesis III*: a natural property is in itself substantial and gives to the object all the substance that it has. These theses are obviously rather different and they tell us not only different, but also contradictory things. *Thesis I*, in fact, is about the independent existence of natural properties and it explicitly refers to the metaphysical distinction between natural and non-natural objects⁸. *Thesis II* is about the role of natural properties as parts of an object. *Thesis III* is about the "nature" or the "quality" of natural properties: according to Moore, in fact, a natural property is a sort of essential property.

Thesis I, then, is not really useful to understand what a natural property is. In this case, in fact, Moore does not probably speak of a natural property as a property of an object, but he rather speaks of that property as a non-natural object. Even if Moore speaks of its existence in time, in fact, when we speak of yellowness, or of size, or of largesse as objects, they are not "object of perception", but they are rather "object of knowledge". *Thesis II*, on the other hand - as Moore himself recognised in 1922 - "involves a confusions between parts and properties of a natural object" (p.13) and then it is not able to allow us to understand the real statute of natural properties. *Thesis III*, finally, is the only one which really tries to define what a natural property is: if a natural property is in fact a substantial property, then it is something like the essential property which Aristotle or Saint Thomas or some other modern and contemporary essentialists (such as Plantinga⁹) have spoken or speaks of. We can then understand why an object could not conceivably lack such a property. In fact, under no possible circumstances an object could fail to possess a property, if this property is an essential one. Nevertheless, this question is a rather difficult one, in that according to essentialists not all properties are essential ones: some properties of an object are in fact essential properties and some others are contingent ones. In this sense, the position of Moore is not yet a clear one, even if the philosopher will analyse the

⁷ PE, p. 93.

⁸ Cf. Moore G.E., *Some Main Problems of Philosophy*, G. Allen & Unwind, London, 1953. This work was written in 1910 as the text of lecture series, but it was left unpublished for over forty years thereafter.

⁹ Cf. Plantinga A., *The Nature of Necessity*, Clarendon, Oxford University Press, Oxford, 1992.

intrinsic/contingent difference in his following works. As in fact he writes in the *Preface* : “Most, if not all, natural and metaphysical predicates, are so far as I can see, either contingent or else intrinsic”¹⁰.

If we now analyse Baldwin’s interpretation of these passages, we can say something more about natural properties. According to Baldwin, in fact, there is only one way to understand Moore’s distinction between natural and non-natural properties: natural properties are independent properties; non-natural properties are dependent properties. As Baldwin writes: “what lies behind Moore’s distinction between natural and non-natural properties is a distinction between independent and dependent properties. Moore contrasts the property of being yellow and being good by saying that something which is yellow has a part which is yellow and nothing else, whereas something which is good lacks a part which is good and nothing else. The contrast he has in mind here is better expressed by the thought that the fact that something is yellow is independent of that thing’s other properties, whereas the fact that something is good is dependent upon its other properties”¹¹. As well as Baldwin, I am convinced that what lies behind Moore’s distinction between natural and non-natural properties is a distinction between independent and dependent properties. Nevertheless, I am also convinced that this distinction cannot really be understood by means of this passage in PE. In order to explain why and how natural properties are independent properties we need in fact to examine some other Moore’s works where the philosopher analyses the relation between intrinsic properties and intrinsic nature of the thing which possesses them. Good, in fact, is not an essential property which constitutes the object and gives to the object all the substance it has. Good is linked to the intrinsic nature of an object and then to its natural properties and in this sense it is a dependent property. But speaking of good as a dependent property does not mean that good could be reduced to a natural property or to some natural properties: one thing is a relation of “dependence”, other thing is a reductionist position according to which good could be reduced to some other properties.

III. Intrinsic properties, contingent properties and intrinsic value

After having analysed this passage in PE, we need an analysis of some others Moore’s works where the philosopher describes the natural/non-natural distinction among properties as the distinction between intrinsic/contingent properties and G. In this sense, Moore does not discard the natural/non-natural distinction - as Baldwin has pointed out - even if he repudiates the account of natural properties given in PE. Moore tries rather better to define this distinction speaking of an intrinsic properties/intrinsic value distinction

¹⁰ Preface, p. 23.

¹¹ Baldwin T., “Ethical Non-Naturalism”, in Hacking (ed.) *Exercises in Analysis*, p. 29.

which is, as he writes in his revised *Preface* "the very same distinction which I should have expressed in *Principia* as the distinction between intrinsic properties which were 'natural' and intrinsic properties which were not 'natural'"¹².

The real problem is now to understand what Moore really means when he speaks of intrinsic properties and intrinsic value. In his revised *Preface* Moore recognises that the most important thing which he wanted to grant about good is that G depends only on the intrinsic nature of states of things which possess it: "So far as I know, the only properties which do not thus depend exclusively on the intrinsic nature of what possesses them, are natural and metaphysical properties"¹³. In certain way, this position can be understood if we remember the interpretation of Baldwin of §26, where he points out the fact that the distinction between natural and non-natural properties is a distinction between independent and dependent properties. G is in fact a property which depends only on the intrinsic nature of the things which possess it and for this reason it is an intrinsic kind of value. On the other hand, natural and metaphysics properties do not depend only on the intrinsic nature of things which possess them and for this reason they must be analysed as contingent, or intrinsic properties.

In this paper I cannot analyse in details the differences between intrinsic and contingent properties. The only thing I can say is that an intrinsic property has a fundamental role in constituting a thing and in defining its intrinsic nature, while a contingent property does not have this same fundamental role and it does not let us immediately understand that A and B are not exactly alike even if A possesses this property and B does not possess it. In fact, if we now call a property ϕ , we can synthesise Moore's position about intrinsic properties in the following way:

if $\phi \in A$

if $\phi \notin B$

and if we immediately understand that $A \neq B$

$\therefore \phi =$ intrinsic property

On the other hand, when we speak of a contingent property we can say that:

if $\phi \in A$

if $\phi \notin B$

and if we don't understand immediately that $A \neq B$

$\therefore \phi =$ contingent property

So, according to Moore, two objects are exactly alike and they have the same intrinsic nature only if their intrinsic properties are the same and if these objects possess these intrinsic properties in the same degree. Intrinsic

¹² *The Philosophy of G. E. Moore* (PGEM), Schilpp P.A. (ed.), Open Court, La Salle, 1968, p. 582 (1st ed. Northwestern, Evanstone, 1942).

¹³ *Preface*, p. 16.

properties are then those the having or lacking of which does not depend on what the rest of the world is like. The intrinsic properties of an object do not depend in fact on what the laws of nature happen to be in the world¹⁴. On the other hand, contingent properties are different from the intrinsic ones because their link with the intrinsic nature of an object that possess them is a different link. The relation between intrinsic nature and intrinsic properties is an internal relation. The relation between intrinsic nature and contingent properties is an external relation. But the claim that objects have some of their properties essentially or necessarily and some other properties accidentally constitute what has been called "essentialism". According to Quine, in fact, essentialism "is the doctrine that some of the attributes of a thing (quite independently of the language in which the thing is referred to, if at all) may be essential to the thing and others accidental"¹⁵.

IV. Essentialism, epistemology and metaphysics

I know it is even suggested in the literature that the notion of a distinction between necessary and contingent properties is just a theory made up by some bad philosophers who didn't realise that there are several ways of referring to the same thing. Nevertheless, this idea that a property can be essential or accidental to an object independently of its description is not an idea which has no intuitive content or which means nothing to the ordinary man. The possible objection is that an ordinary man does not have a philosophical theory in order to justify his intuition. But the possible answer is that there are in the world different ways of being philosophers. In this sense we do not only have philosophers who think that it is only the description which makes a property an essential property of an object, but we have also philosophers who think that things may be said to have modal properties and who assume that it makes sense to speak of essentialism. Such an assumption is, of course, controversial, but it may be of interest, as well as it may be of interest the choice of adopting a rich ontology instead of a poor (and sometimes sterile) one.

¹⁴ Cf. the position of Vallentyne ("Intrinsic Properties Defined", *Philosophical Studies*, 88, 1997, pp. 209-219) who develops this position and introduces the notion of "contraction of a world" according to which the world considered is a world obtained by removing some of the objects in the original world. In particular this contraction is called by Vallentyne a "x-t-contraction" (where x is an object and t is a time): "The intuitive idea is that such a contraction is a world 'obtainable' from the original one by, to the greatest extent possible, 'removing' all objects wholly distinct from x, all spatial locations not occupied by x, and all times (temporal states of the world) except t, from the world" (p. 211). The result of this x-t-contraction is then a lawless world in which we can imagine an exact duplicate of an object which is now in a world with different laws of nature.

¹⁵ Quine W., "Three Grades of Modal Involvement", *The Ways of Paradox*, Random House, New York, 1966, p.173.

I know that some critics have tried to show that the position of Moore can be defended from the attacks of his critics if we show that his philosophy is no more an ontological and metaphysical one, but rather a logical position. According to Baldwin, for example, the only way to defend Moore's ethical non-naturalism is to analyse the distinction between natural and non-natural properties as a logical distinction rather than as an ontological one. In this way, in fact, "some of the familiar criticisms of Moore's ethical non-naturalism (such as Mackie's Argument from Queerness, 1977) turn out to attend to Moore's rhetoric, but not to his argument"¹⁶. Nevertheless, I am convinced that we can "defend" Moore's position also if we try to understand his position, even if his position is an ontological one. In this sense I do not want to try to defend essentialism, also because I am not able to do this. The only thing I can say, but that I probably cannot demonstrate, is that a poor ontology is not always the best way to understand the world and to approach some main problems of philosophy. In this sense, I do not see the force of many criticism against a rich ontology, such as that of Quine. Even if from an epistemological point of view Quine keeps some sympathy, one thing is an epistemological approach to the world, other thing is a metaphysical approach, as Kripke¹⁷ has rightly pointed out. So, even if from an epistemological point of view Moore's position is not always able to convince, from an ontological point of view it is a rather interesting one and it can let us better understand not only the intrinsic properties/intrinsic nature relation but also the intrinsic value/intrinsic nature relation.

V. The relation between intrinsic properties, intrinsic nature and intrinsic value

After having analysed the intrinsic properties/intrinsic value distinction, in order better to understand the role of G we have to analyse either the G/intrinsic nature relation (G/N_i relation), or the intrinsic properties/intrinsic nature relation (P_i/N_i relation). If we analyse the paper CIV, we can easily understand that both these relations are necessary relations. Speaking in fact of the necessity that links together the intrinsic value of a thing and its intrinsic nature, Moore points out that this necessity is not a causal, nor an empirical, nor a logical necessity. Then, in order to explain this sort of "ontological necessity", he analyses the example of a particular patch of colour which is yellow and underlines that any patch of colour exactly like that one considered would be yellow "even if it existed in a Universe in which causal law were quite different from what they are in this one"¹⁸. In this sense, according to

¹⁶ Baldwin T., "Ethical Non-Naturalism", in Hacking (ed.) *Exercises in Analysis*, p. 30.

¹⁷ Kripke S., *Naming and Necessity*, Blackwell, London, 1980.

¹⁸ CIV, p. 292.

Moore “any such patch must be yellow, quite unconditionally, whatever the circumstances and whatever the causal laws”¹⁹.

But which is then the difference? Why G is different from an intrinsic property and it can never be reduced to it? Even if intrinsic properties are completely different from G because G is a dependent property while intrinsic properties are independent properties, both the G/N_i relation and the P_i/N_i relation are “necessary relations”: they are both relations where there is the necessity that, if x and y possess different intrinsic properties or different intrinsic value, their nature must be different. Nevertheless, the necessity involved in these two relations is a rather different one. The necessity that, if x and y possess different intrinsic properties, their nature must be different “is a necessity of a different kind from the necessity there is that, if x and y are of different intrinsic value, their nature must be different, although both necessities are unconditional”²⁰.

When speaking about the necessary relation between intrinsic nature and intrinsic value, Moore intends that a kind of value is intrinsic if and only if, when anything possesses it, the same thing or anything exactly like it would necessary or must always possess it in exactly the same degree. If a kind of value is intrinsic, it must be impossible for things to possess it in different degrees, if they are exactly like one another. Nevertheless, even if intrinsic value depends on the intrinsic nature, intrinsic nature does not depend on the intrinsic value. If a thing is good, that follows from the fact that it possesses certain intrinsic properties, but “from the fact that it is good it does not follow conversely that it has those properties”²¹. An intrinsic value, in fact, never describes the intrinsic nature of what possesses it, nor it has to be mentioned when we give a complete description of a thing. On the contrary, the intrinsic properties describe the intrinsic nature of what possesses them and they must be mentioned in a complete description of a thing: “intrinsic properties seem to *describe* the intrinsic nature of what possesses them in a sense in which predicates of value never do. If you could enumerate *all* the intrinsic properties a given thing possessed, you would have given a *complete* description of it, and would not need to mention any predicates of value it possessed; whereas no description of a given thing could be *complete* which omitted any intrinsic property”²².

In this sense the P_i/N_i relation is completely from the G/N_i relation. When we speak of the relation between intrinsic nature and intrinsic properties, we have to recognise that the intrinsic nature depends on the intrinsic properties because these properties are able to describe the intrinsic

¹⁹ CIV, p. 293.

²⁰ CIV, p. 298.

²¹ PGEM, p. 588.

²² CIV, p. 297.

nature of a thing: if the intrinsic properties of the thing A are different from the intrinsic properties of the thing B, then their intrinsic nature is different. On the contrary, when we speak of the relation between intrinsic nature and intrinsic value, we have to recognise that the intrinsic value depends on the intrinsic nature of a thing: if a thing A is good, from the fact that it is good it does not follow that it has the same intrinsic properties of the thing B, even if B is good, because good is not a property which can describe the intrinsic nature of A, or the intrinsic nature of B.

For this reason the G/N_i relation looks like this: (1) $G = f(N_i)$. But this relation never can be expressed as: (2) $N_i = f(G)$.

On the other hand, the P_i/N_i relation looks like this: (3) $N_i = f(P_i)$. And it never can be expressed as: (4) $P_i = f(N_i)$.

Good depends on the intrinsic nature of a thing that possesses it in that it is directly linked to the intrinsic nature; G is always "function" of the intrinsic nature and then, G is the dependent variable, while the intrinsic nature (N_i) is the independent variable. On the other hand, intrinsic properties are not "function" of the intrinsic nature and in this sense they are independent properties. Then, N_i is the dependent variable, while P_i is the independent variable, even if there is a relation between intrinsic nature and intrinsic properties in that it is the intrinsic nature which depends on the intrinsic properties.

The relations between intrinsic properties and intrinsic nature of a thing is a relation of dependence because the nature of a thing depends on these properties and in this sense this position is not really far from that in PE when Moore, speaking of natural properties, writes that these properties "are in themselves substantial and give to the object all the substance I has". The difference is that now not all natural properties are according to Moore substantial ones, because only the intrinsic ones give all the substance to the objects and are then necessities to the nature of this thing.

If we substitute (3) in (1) we have then:

$$(5) G = f(f(P_i))$$

Finally, if $f \circ f = g$ we arrive at:

$$(6) G = g(P_i)$$

The relation between G and the intrinsic properties of a thing is then a dependence relation, but G can never be completely reduced to the intrinsic properties of the thing that possesses it. One should distinguish between "dependence" and "reduction". The assumption that once there are connections between intrinsic value and intrinsic properties, it must be possible to represent G as reducible to the intrinsic properties of an object is unjustified. According to Moore, in fact, G is different from intrinsic properties either because G is not able to describe the intrinsic nature of an object, or because G can never be considered an essential property. As we can

read in Moore's reply to his critics in PEGM, in fact: "I should never have thought of suggesting that goodness was 'non-natural', unless I had supposed that it was 'derivative' in the sense that, whenever a thing is good its goodness (in Mr Broad's words) 'depends on the presence of certain non-ethical characteristics' possessed by the thing in question: I have always supposed that it did so 'dependent', in the sense that, if a thing is good (in may sense), then that this is so follows from the fact that it has certain natural intrinsic properties, which are such that from the fact that it is good it does not follow conversely that it has those properties"²³.

The real problem now is to understand what is exactly G in "positive terms". We know in fact that it is not a substantial or essential property, that it is a dependent property and that it can never be reduced to some other intrinsic properties. On the other hand we know that its relation with the intrinsic nature of a thing is a necessary relation, even if it is completely different from the P_i/N_i relation. But then how could we "define" G? In PE Moore writes that "good is good and that is the end of the matter". Nevertheless, my hypothesis is that from an ontological point of view G may be considered as a "transcendental" in the medieval sense of the expression, as the Italian philosopher Preti²⁴ has once pointed out. But when saying that good is a transcendental I probably begin another chapter in the analysis of Moore's position and then I lay the foundations of a future work.

REFERENCES

- Baldwin T., "Ethical Non-Naturalism", in Hacking I. (ed.) *Exercises in Analysis*, Cambridge University Press, Cambridge, 1985, pp. 24-45
 —, *G.E. Moore*, Routledge, London/New York, 1990
 Humberstone I.L., "Intrinsic/Extrinsic", *Synthese*, 108, 1996, pp. 205-267
 Kripke S., *Naming and Necessity*, Blackwell, London, 1980
 Lewis D., "Extrinsic Properties", *Philosophical Studies*, 44, 1983, pp. 197-200
 —, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61, 1983, pp. 343-377
 Moore G.E., *Principia Ethica*, Cambridge University Press, Cambridge, 1993
 —, "The Conception of Intrinsic Value", in *Philosophical Studies*, Routledge, London, 1922
 —, *Some Main Problems of Philosophy*, G. Allen & Unwind, London, 1953
 —, "Reply to my Critics", *The Philosophy of G. E. Moore*, Schilpp P.A. (ed.), Open Court, La Salle, 1968
 O' Connor D., *The metaphysics of G. E. Moore*, Reidel Company, Boston/London, 1982
 Plantinga A., *The Nature of Necessity*, Clarendon, Oxford University Press, Oxford, 1992
 Preti G., *Il problema dei valori: l'etica di G.E. Moore*, Angeli, Milano, 1986
 Quine W., "Three Grades of Modal Involvement", *The Ways of Paradox*, Random House, New York, 1966
 Strawson P.E., *Individuals. An Essay in Descriptive Metaphysics*, Routledge, London-New York, 1959

²³ PGEM, p. 588.

²⁴ Preti G., *Il problema dei valori: l'etica di G.E. Moore*, Angeli, Milano, 1986.

Sylvester P., *The Moral Philosophy of G. E. Moore*, Temple University Press, Philadelphia, 1990

Vallentyne P., "Intrinsic Properties Defined", *Philosophical Studies*, 88, 1997, pp. 209-219

Maria Michela Marzano-Parisoli

Centre de recherche Sens, éthique et société - CNRS

e-mail: luca.parisoli@u-paris10.fr



C U R S O S E C O N G R E S O S
Nº 120



XACOBEO'99



CONSELLERÍA DE CULTURA,
COMUNICACIÓN SOCIAL E TURISMO

CONSELLERÍA DA PRESIDENCIA
SECRETARÍA XERAL DE INVESTIGACIÓN
E DESENVOLVEMENTO



MINISTERIO DE EDUCACIÓN Y CULTURA



Residencia Universitaria
San Martín Pinario

