Studies in Applied Philosophy, Epistemology and Rational Ethics

SAPERE

Ángel Nepomuceno-Fernández
Lorenzo Magnani
Francisco J. Salguero-Lamillar
Cristina Barés-Gómez
Matthieu Fontaine *Editors*

Model-Based Reasoning in Science and Technology

Inferential Models for Logic, Language, Cognition and Computation

Springer

Studies in Applied Philosophy, Epistemology and Rational Ethics

Volume 49

Editor-in-Chief

Lorenzo Magnani, Department of Humanities, Philosophy Section, University of Pavia, Pavia, Italy

Editorial Board Members

Atocha Aliseda
Universidad Nacional Autónoma de México (UNAM), Mexico, Mexico

Giuseppe Longo
CNRS - Ecole Normale Supérieure, Centre Cavailles, Paris, France

Chris Sinha
School of Foreign Languages, Hunan University, Changsha, China

Paul Thagard
University of Waterloo, Waterloo, Canada

John Woods
University of British Columbia, Vancouver, Canada

Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE) publishes new developments and advances in all the fields of philosophy, epistemology, and ethics, bringing them together with a cluster of scientific disciplines and technological outcomes: ranging from computer science to life sciences, from economics, law, and education to engineering, logic, and mathematics, from medicine to physics, human sciences, and politics. The series aims at covering all the challenging philosophical and ethical themes of contemporary society, making them appropriately applicable to contemporary theoretical and practical problems, impasses, controversies, and conflicts. Our scientific and technological era has offered "new" topics to all areas of philosophy and ethics – for instance concerning scientific rationality, creativity, human and artificial intelligence, social and folk epistemology, ordinary reasoning, cognitive niches and cultural evolution, ecological crisis, ecologically situated rationality, consciousness, freedom and responsibility, human identity and uniqueness, cooperation, altruism, intersubjectivity and empathy, spirituality, violence. The impact of such topics has been mainly undermined by contemporary cultural settings, whereas they should increase the demand of interdisciplinary applied knowledge and fresh and original understanding. In turn, traditional philosophical and ethical themes have been profoundly affected and transformed as well: they should be further examined as embedded and applied within their scientific and technological environments so to update their received and often old-fashioned disciplinary treatment and appeal. Applying philosophy individuates therefore a new research commitment for the 21st century, focused on the main problems of recent methodological, logical, epistemological, and cognitive aspects of modeling activities employed both in intellectual and scientific discovery, and in technological innovation, including the computational tools intertwined with such practices, to understand them in a wide and integrated perspective.

Studies in Applied Philosophy, Epistemology and Rational Ethics means to demonstrate the contemporary practical relevance of this novel philosophical approach and thus to provide a home for monographs, lecture notes, selected contributions from specialized conferences and workshops as well as selected PhD theses. The series welcomes contributions from philosophers as well as from scientists, engineers, and intellectuals interested in showing how applying philosophy can increase knowledge about our current world. Initial proposals can be sent to the Editor-in-Chief, prof. Lorenzo Magnani, lmagnani@unipv.it:

- A short synopsis of the work or the introduction chapter
- The proposed Table of Contents
- The CV of the lead author(s)

For more information, please contact the Editor-in-Chief at lmagnani@unipv.it.

Indexed by SCOPUS, ISI and Springerlink. The books of the series are submitted for indexing to Web of Science.

More information about this series at http://www.springer.com/series/10087

Ángel Nepomuceno-Fernández ·
Lorenzo Magnani ·
Francisco J. Salguero-Lamillar ·
Cristina Barés-Gómez · Matthieu Fontaine
Editors

Model-Based Reasoning in Science and Technology

Inferential Models for Logic, Language, Cognition and Computation

Springer

Editors
Ángel Nepomuceno-Fernández
Department of Philosophy,
Logic and Philosophy of Science
University of Seville
Seville, Spain

Francisco J. Salguero-Lamillar
Department of Spanish Language,
Linguistics and Theory of Literature
University of Seville
Seville, Spain

Matthieu Fontaine
Centre for Philosophy of Science of the
University of Lisbon
University of Lisbon
Lisbon, Portugal

Lorenzo Magnani
Department of Humanities,
Philosophy Section, and Computational
Philosophy Laboratory
University of Pavia
Pavia, Italy

Cristina Barés-Gómez
Department of Philosophy,
Logic and Philosophy of Science
University of Seville
Seville, Spain

ISSN 2192-6255 ISSN 2192-6263 (electronic)
Studies in Applied Philosophy, Epistemology and Rational Ethics
ISBN 978-3-030-32721-7 ISBN 978-3-030-32722-4 (eBook)
https://doi.org/10.1007/978-3-030-32722-4

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume is a collection of selected papers that were presented at the International Conference *Model-Based Reasoning in Science and Technology. Inferential Models for Logic, Language, Cognition and Computation* (MBR018_SPAIN), held at Tobacco Factory University Building, Seville, Spain, October 24–26, 2018, chaired by Ángel Nepomuceno, Lorenzo Magnani, and Francisco J. Salguero.

This event marked the twentieth anniversary of the model-based reasoning conferences, since the first meeting was held at the Collegio Ghislieri, University of Pavia, Pavia, Italy, in December 1998.

A previous volume, *Model-Based Reasoning in Scientific Discovery*, edited by L. Magnani, N. J. Nersessian, and P. Thagard (Kluwer Academic/Plenum Publishers, New York, 1999; Chinese edition, China Science and Technology Press, Beijing, 2000), was based on the papers presented at the first "model-based reasoning" international conference, held at the University of Pavia, Pavia, Italy, in December 1998. Other two volumes were based on the papers presented at the second "model-based reasoning" international conference, held at the same place in May 2001: *Model-Based Reasoning. Scientific Discovery, Technological Innovation, Values*, edited by L. Magnani and N. J. Nersessian (Kluwer Academic/Plenum Publishers, New York, 2002), and *Logical and Computational Aspects of Model-Based Reasoning*, edited by L. Magnani, N. J. Nersessian, and C. Pizzi (Kluwer Academic, Dordrecht, 2002). Another volume, *Model-Based Reasoning in Science and Engineering*, edited by L. Magnani (College Publications, London, 2006), was based on the papers presented at the third "model-based reasoning" international conference, held at the same place in December 2004. The volume *Model-Based Reasoning in Science and Medicine*, edited by L. Magnani and L. Ping (Springer, Heidelberg/Berlin 2006), was based on the papers presented at the fourth "model-based reasoning" conference, held at Sun Yat-sen University, Guangzhou, P. R. China. The volume *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*, edited by L. Magnani, W. Carnielli, and C. Pizzi (Springer, Heidelberg/Berlin 2010), was based on the papers presented at the fifth "model-based reasoning"

conference, held at the University of Campinas, Campinas, Brazil, in December 2009. The volume *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, edited by L. Magnani, (Springer, Heidelberg/Berlin 2013), was based on the papers presented at the sixth "model-based reasoning" conference, held at Fondazione Mediaterraneo, Sestri Levante, Italy, in June 2012. Finally, the volume *Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues*, edited by L. Magnani and C. Casadio (Springer, Cham, Switzerland 2016), was based on the papers presented at the seventh "model-based reasoning" conference, held at Centro Congressi Mediaterraneo, Sestri Levante, Italy, in June 2015.

The presentations given at the Seville conference explored how scientific thinking uses models and explanatory reasoning to produce creative changes in theories and concepts. Some speakers addressed the problem of model-based reasoning in technology and stressed issues such as the relationship between science and technological innovation. The study of diagnostic, visual, spatial, analogical, and temporal reasoning has demonstrated that there are many ways of performing intelligent and creative reasoning that cannot be described with the help only of traditional notions of reasoning such as classical logic. Understanding the contribution of modeling practices to discovery and conceptual change in science and in other disciplines requires expanding the concept of reasoning to include complex forms of creativity that are not always successful and can lead to incorrect solutions.

The study of these heuristic ways of reasoning is situated at the crossroads of philosophy, artificial intelligence, cognitive psychology, and logic, i.e., at the heart of cognitive science. There are several key ingredients common to the various forms of model-based reasoning. The term "model" comprises both internal and external representations. The models are intended as interpretations of target physical systems, processes, phenomena, or situations. The models are retrieved or constructed on the basis of potentially satisfying salient constraints of the target domain. Moreover, in the modeling process, various forms of abstraction are used. Evaluation and adaptation take place in light of structural, causal, and/or functional constraints. Model simulation can be used to produce new states and enable evaluation of behaviors and other factors. The various contributions of the book are written by interdisciplinary researchers who are active in the area of modeling reasoning and creative reasoning in logic, cognitive science, science, and technology: The most recent results and achievements about the topics above are illustrated in detail in the papers.

The editor expresses his appreciation to the members of the Scientific Committee for their suggestions and assistance: Selene Arfini, Computational Philosophy Laboratory, Department of Humanities, Philosophy Section, University of Pavia, Italy; Atocha Aliseda, Instituto de Investigaciones Filosoficas, Universidad Nacional Autonoma de Mexico (UNAM); Francesco Amigoni, Politecnico di Milano, Dipartimento di Elettronica, Informazione, e Bioingegneria, Milano, Italy; Tommaso Bertolotti, Department of Humanities, Philosophy Section, University of Pavia, Italy; Otávio Bueno, Department of Philosophy, University of Miami, Coral

Gables, USA; Walter Carnielli, Department of Philosophy, Institute of Philosophy and Human Sciences, State University of Campinas, Brazil; Claudia Casadio, Department of Philosophy, Education and Economical-Quantitative Sciences, University of Chieti-Pescara, Italy; Sanjay Chandrasekharan, Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research, India; Sara Dellantonio, Psychology and Cognitive Sciences, University of Trento, Italy; Gordana Dodig-Crnkovic, Chalmers University of Technology, Department of Applied Information Technology, Göteborg, Sweden; Maria Gulia Dondero, Maître de recherches du FNRS, Université de Liège, Belgium; Steven French, Department of Philosophy, University of Leeds, Leeds, UK; Roman Frigg, London School of Economics and Political Science, UK; Marcello Frixione, Department of Communication Sciences, University of Salerno, Italy; Dov Gabbay, Department of Computer Science, King's College, London, UK; Axel Gelfert, Professor of Philosophy, Technical University of Berlin, Germany; Valeria Giardino, Archives Henri-Poincaré UMR 7117 CNRS–Université de Lorraine, Nancy, FRANCE Marcello Guarini, Department of Philosophy, University of Windsor, Canada; Ricardo Gudwin, Department of Computer Engineering and Industrial Automation, the School of Electrical Engineering and Computer Science, State University of Campinas, Brazil; Albrecht Heeffer, Centre for History of Science, Ghent University, Belgium; Decio Krause, Departamento de Filosofia, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil; Ping Li, Department of Philosophy, Sun Yat-sen University, Guangzhou, P.R. China; Angelo Loula, Department of Exact Sciences, State University of Feira de Santana, Brazil; Lorenzo Magnani, Department of Humanities, Philosophy Section and Computational Philosophy Laboratory, University of Pavia, Italy; Joke Meheus, Vakgroep Wijsbegeerte, Universiteit Gent, Gent, Belgium; Luís Moniz Pereira, Departamento de Informática, Universidade Nova de Lisboa, Portugal; Michael Moortgat, Utrecht University, Institute of Linguistics (OTS), Utrecht, The Netherlands; Woosuk Park, Humanities and Social Sciences, KAIST, Guseong-dong, Yuseong-gu Daejeon, South Korea; Mario J. Pérez, Department of Computer Sciences, Academy of Europe, University of Seville, Spain; Ahti-Veikko Pietarinen, Ragnar Nurkse School of Innovation and Governance, Tallin University of Technology, Estonia, and School of Humanities and Social Sciences, Nazarbayev University, Kazakhstan; Claudio Pizzi, Department of Philosophy and Social Sciences, University of Siena, Siena, Italy; Olga Pombo, Centro de Filosofia das Ciências Universidade de Lisboa (CFCUL), Portugal; Demetris Portides, Department of Classics and Philosophy, University of Cyprus, Nicosia, Cyprus; Joao Queiroz, Institute of Arts and Design, Federal University of Juiz de Fora, Brazil; Shahid Rahman, UFR de Philosophie, University of Lille 3, Villeneuve d'Ascq, France; Oliver Ray, Department of Computer Science, University of Bristol, Bristol, UK; Flavia Santoian, Dipartimento Studi Umanistici, Universitá di Napoli Federico II, Italy; Colin Schmidt, Le Mans University and ENSAM-ParisTech, France; Gerhard Schurz, Institute for Philosophy, Heinrich-Heine University, Frankfurt, Germany; Nora Alejandrina Schwartz, Faculty of Economics, Universidad de Buenos Aires, Argentina; Cameron Shelley,

Department of Philosophy, University of Waterloo, Waterloo, Canada; Sonja Smets, Institute for Logic, Language and Computation (ILLC), University of Amsterdam, The Netherlands; Nik Swoboda, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain; Adam Toon, Sociology, Philosophy and Anthropology, University of Exeter, UK; Paul Thagard, Department of Philosophy, University of Waterloo, Waterloo, Canada; Barbara Tversky, Department of Psychology, Stanford University and Teachers College, Columbia University, New York, USA; Ryan D. Tweney, Emeritus Professor of Psychology, Bowling Green State University, Bowling Green, USA; Hans van Ditmarsch, Loria, Nancy, France; Fernando Velazquez, Institute for Logic, Language and Computation (ILLC), University of Amsterdam, The Netherlands; Riccardo Viale, Scuola Nazionale dell'Amministrazione, Presidenza del Consiglio dei Ministri, Roma, and Fondazione Rosselli, Torino, Italy; John Woods, Department of Philosophy, University of British Columbia, Canada.

We are also very thankful to the local scientific committee: Cristina Barés, Logic, Philosophy and Philosophy of Science Department, University of Seville, Spain; Alfredo Burrieza, Department of Philosophy, University of Málaga, Spain; Matthieu Fontaine, CFCUL, Lisbon University, Portugal; Teresa López-Soto, English Language Department, University of Seville, Spain; Angel Nepomuceno, Logic, Philosophy and Philosophy of Science Department, University of Seville, Spain; José F. Quesada, Computer Sciences and Artificial Intelligence Department, University of Seville, Spain; Francisco J. Salguero, Department of Spanish Language, Linguistics and Theory of Literature, University of Seville, Spain; Fernando Soler, Department of Logic, Philosophy and Philosophy of Science, University of Seville, Spain. We warmly thank local organizers for their help: Nino Guallart, Rocio Ramírez, Pablo Sierra, Logic, Philosophy and Philosophy of Science Department, University of Seville, Spain; Diego Jímenez Department of Spanish Language, Linguistics and Theory of Literature, University of Seville, Spain.

The conference MBR018_SPAIN, and thus indirectly this book, was made possible through the generous financial support of the Italian Ministry of the University (MIUR) and of the University of Pavia. Their support is gratefully acknowledged. The preparation of the volume would not have been possible without the contribution of resources and facilities of the Computational Philosophy Laboratory and of the Department of Humanities, Philosophy Section, University of Pavia, and of the University of Seville, the Faculties of Philology and Philosophy, the Department of Philosophy, Logic and Philosophy of Science, and the Research Group on Logic, Language and Information of the University of Seville.

Several papers concerning model-based reasoning deriving from the previous conferences MBR98 and MBR01 can be found in special issues of journals: in *Philosophica*: Abduction and Scientific Discovery, 61(1), 1998, and Analogy and Mental Modeling in Scientific Discovery, 61(2) 1998; in *Foundations of Science*: Model-Based Reasoning in Science: Learning and Discovery, 5(2) 2000, all edited by L. Magnani, N. J. Nersessian, and P. Thagard; in *Foundations of Science*: Abductive Reasoning in Science, 9, 2004, and Model-Based Reasoning: Visual,

Analogical, Simulative, 10, 2005; in *Mind and Society*: Scientific Discovery: Model-Based Reasoning, 5(3), 2002, and Commonsense and Scientific Reasoning, 4(2), 2001, all edited by L. Magnani and N. J. Nersessian. Finally, other related philosophical, epistemological, and cognitive-oriented papers deriving from the presentations given at the conference MBR04 have been published in a Special Issue of the *Logic Journal of the IGPL*: Abduction, Practical Reasoning, and Creative Inferences in Science, 14(1) (2006) and have been published in two Special Issues of *Foundations of Science*: Tracking Irrational Sets: Science, Technology, Ethics, and Model-Based Reasoning in Science and Engineering, 13(1) and 13(2) (2008), all edited by L. Magnani. Other technical logical papers presented at MBR09_BRAZIL have been published in a special issue of the *Logic Journal of the IGPL*: Formal Representations in Model-Based Reasoning and Abduction, 20(2) (2012), edited by L. Magnani, W. Carnielli, and C. Pizzi. Then, technical logical papers presented at MBR12_ITALY have been published in a special issue of the *Logic Journal of the IGPL*: Formal Representations in Model-Based Reasoning and Abduction, 21(6) (2013), edited by L. Magnani. Finally, technical papers presented at MBR15_ITALY have been published in a special issue of the *Logic Journal of the IGPL*: Formal representations of model-based reasoning and abduction, 24(4) (2016), edited by L. Magnani and C. Casadio.

Other more technical formal papers presented at (MBR18_SPAIN) will be published in a special issue of the *Logic Journal of the IGPL*: Formal representations of model-based reasoning and abduction, edited by A. Nepomuceno, L. Magnani, F. Salguero, C. Barés, and M. Fontaine.

July 2019

Ángel Nepomuceno-Fernández
Lorenzo Magnani
Francisco J. Salguero-Lamillar
Cristina Barés-Gómez
Matthieu Fontaine

Contents

Models, Mental Models, and Representations

Probing Possibilities: Toy Models, Minimal Models, and Exploratory Models 3
Axel Gelfert

Model Types and Explanatory Styles in Cognitive Theories 20
Simone Pinna and Marco Giunti

The Logic of Dangerous Models 41
Selene Arfini

A Pragmatic Model of Justification Based on "Material Inference" for Social Epistemology 55
Raffaela Giovagnoli

Counterfactual Thinking in Cooperation Dynamics 69
Luís Moniz Pereira and Francisco C. Santos

Modeling Morality 83
Walter Veit

Coherence and Credibility in the Story-Model of Jurors' Decision-Making: Does Mental Simulation Really Drive the Evaluation of the Evidence? 103
Marion Vorms and David Lagnado

Insight Problem Solving and *Unconscious Analytic* Thought. New Lines of Research 120
Laura Macchi, Veronica Cucchiarini, Laura Caravona, and Maria Bagassi

On Understanding and Modeling in Evo-Devo 138
Rodrigo Lopez-Orellana and David Cortés-García

Conjuring Cognitive Structures: Towards a Unified Model of Cognition 153
Majid D. Beni

How Philosophical Reasoning and Neuroscientific Modeling Come Together 173
Gabriele Ferretti and Marco Viola

Abduction, Problem Solving, and Practical Reasoning

The Dialogic Nature of Semiotic Tools in Facilitating Conscious Thought: Peirce's and Vygotskii's Models 193
Donna E. West

Creative Model-Based Diagrammatic Cognition 217
Lorenzo Magnani

Kant on the Generality of Model-Based Reasoning in Geometry 245
William Goodwin

The Logic of Picturing: Wittgenstein, Sellars and Peirce's EG-beta 256
Rocco Gangle, Gianluca Caterina, and Fernando Tohmé

An Inferential View on Human Intuition and Expertise 274
Rico Hermkes and Hanna Mach

Disseminated Causation: A Model-Theoretical Approach to Sophisticated Abduction 287
Andrés Rivadulla

Defining a General Structure of Four Inferential Processes by Means of Four Pairs of Choices Concerning Two Basic Dichotomies 298
Antonino Drago

Remarks on the Possibility of Ethical Reasoning in an Artificial Intelligence System by Means of Abductive Models 318
Alger Sans and David Casacuberta

Epistemological and Technological Issues

On the Follies of Intercourse Between Models and Fiction: A Naturalized Causal-Response Diagnosis 337
John Woods

Default Soundness in the Old Approach: An Epistemic Analysis of Default Reasoning 372
David Gaytán

Models and Data in Finance: *les Liaisons Dangereuses* 393
Emiliano Ippoliti

How Can You Be Sure? Epistemic Feelings as a Monitoring System for Cognitive Contents 407
Sara Dellantonio and Luigi Pastore

A Model for the Interlock Between Propositional and Motor Formats 427
Gabriele Ferretti and Silvano Zipoli Caiani

A Computational-Hermeneutic Approach for Conceptual Explicitation 441
David Fuenmayor and Christoph Benzmüller

The Context-Priming of Conceptual Knowledge: RPEC and SPEC Models 470
Bideng Xiang and Ping Li

Graphs in Linguistics: Diagrammatic Features and Data Models 482
Paolo Petricca

Author Index 501

Models, Mental Models, and Representations



Probing Possibilities: Toy Models, Minimal Models, and Exploratory Models

Axel Gelfert(✉)

Chair of Theoretical Philosophy, Technische Universität Berlin,
Straße des 17. Juni 135, H72, 10623 Berlin, Germany
axel@gelfert.net

Abstract. According to one influential view, model-building in science is primarily a matter of simplifying theoretical descriptions of real-world target systems using abstraction and idealization. This view, however, does not adequately capture all types of models. Many contemporary models in the natural and social sciences – from physics to biology to economics – stand in a more tenuous relationship with real-world target systems and have a decidedly stipulative element, in that they create, by fiat, 'model worlds' that operate according to some set of specified rules. While such models may be motivated by an interest in actual target phenomena, their validity is not – at least not primarily – to be judged by whether they constitute an empirically adequate representation of any particular empirical system. The present paper compares and contrasts three such types of models: minimal models, toy models, and exploratory models. All three share some characteristics and thus overlap in interesting ways, yet they also exhibit significant differences. It is argued that, in all three cases, modal considerations have an important role to play: by exploring the modal structure of theories and phenomena – that is, by probing possibilities in various ways – such models deepen our understanding and help us gain knowledge not only about what there is in the world, but also about what there could be.

Keywords: Scientific models · Exploratory models · Minimal models · Toy models

1 Introduction

Scientific models, according to one important line of philosophical analysis, aim at representing real-world target systems, even if they only ever do so imperfectly. Even where an underlying theory is available, a full description of a target is often out of reach, so simplified models need to be derived using abstraction and idealization. On this analysis, the resulting models are to be assessed as representations of actual target systems, and the success of a model is judged by how closely it resembles the (elusive) full description of the target system. Yet, whatever initial plausibility this account of the practice of scientific modelling might have, it does by no means adequately describe all types of modelling. In particular, many contemporary models across the natural and social sciences – from physics to biology to economics – stand in a more tenuous

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 3–19, 2019.
https://doi.org/10.1007/978-3-030-32722-4_1

relationship with real-world target systems, and deliberately so. The present paper discusses three such types of models: minimal models, toy models, and exploratory models. Minimal models, which have received some philosophical attention in connection with the physics of phase transitions, have been variously described as "thoroughgoing caricatures of real systems" or even as "really look[ing] nothing like any system [they are] supposed to 'represent'" (Batterman and Rice 2014). Toy models – that is, models so idealized and simplified that they border on being 'stylized' accounts of a single aspect of a (real or hypothetical) phenomenon – have received conflicting interpretations: whereas some authors deny outright that toy models can serve a representational function, others insist that they do. Exploratory models, finally, are employed whenever an integrated body of theoretical knowledge cannot be assumed (either because such knowledge is itself a matter of dispute or because the subject matter does not allow for an underlying 'fundamental' theory). Yet, the utility and continued use of such models calls out for an explanation. Why do such models persist? The present paper agues that the key to this question lies in recognizing that models frequently explore the modal structure of theories and phenomena; that is, they help us understand what is, and isn't, possible within a certain segment of the real world. Models are ways of probing possibilities as much as they are (sometimes) representations of real-world target systems.

The rest of this paper is structured as follows. Section 2 explores a persistent tension in philosophical accounts of scientific modelling, which is marked by, on the one hand, *demanding too much* from scientific models (e.g. truth as a precondition for explanatory success) while, on the other hand, *underestimating* what scientific models can already achieve (e.g., providing modal insight across a range of possible scenarios). Section 3 discusses how scientific practice, including the practice of modelling, is guided by competing regulative ideals, such as completeness and simplicity, which trade off against each other. Section 4 extends this discussion to different types of idealization, which can partly be seen as ways of implementing those representational ideals. Section 5 compares and contrasts, in turn, minimal models, toy models, and exploratory models. For each type of model, definitions found in the philosophical literature are presented and discussed, along with illustrative examples. Special attention is devoted to the import of *modal considerations* in all three cases. The paper ends with a brief section outlining further lines of inquiry and concludes that closer engagement with the history of scientific modelling may be able to cast light on the continuities, and discontinuities, among the various types of models across different phases in the process of scientific inquiry.

2 Explanation and Modelling: An Essential Tension

On a popular construal of science, it is in the business of giving explanations. There may be all sorts of other expectations we have of science – that it should give us an accurate description of what there is in the world, that it should bring about technological applications that improve our lot – but, for theoretical science at least, an influential view is that, collectively, it should allow us to progressively satisfy our curiosity about the workings of the world. Arguably, scientific modelling is one of the

key components of the methodological arsenal of science, standing alongside experimentation, observation, and scientific theorizing. While, historically, it took philosophers of science some time to acknowledge the centrality of models to scientific practice, over the past twenty-five years or so a consensus seems to have emerged that modelling is an indispensable part of the methodological toolbox of science.

And yet, there seems to be a lingering tension between the high demands of scientific explanation, on the one hand, and the inevitable – and widely acknowledged – limitations of scientific modelling, on the other hand. As a result of this tension, philosophical debates about scientific models have tended to oscillate between two deeply rooted 'gut reactions' which are likewise in tension with one another: namely, on the one hand, *demanding too much* from scientific models, while simultaneously, on the other hand, *underestimating* what scientific models *can* achieve. The tension is roughly similar to what Julian Reiss, in regard to economic models, has dubbed "the explanation paradox" (Reiss 2012). The purported paradox – which Reiss claims has not yet been decisively resolved – arises from the following inconsistent triad: "(1) Economic models are false. (2) Economic models are nevertheless explanatory. (3) Only true accounts can explain." (Reiss 2012, p. 49) In other words, while economic models distort reality (often deliberately so), they are nonetheless being credited with doing explanatory work – in spite of the fact that our best philosophical theories of scientific explanation require faithful representation as a condition for explanatory success.

The faithfulness requirement takes different forms according to one's preferred philosophical account of explanations. In its starkest form, it amounts to no less than the demand for full-blown truth. Famously, or notoriously, the truth of the explanans is one of the key conditions of adequacy of Hempel and Oppenheim's deductive-nomological account of explanation. That is, for a model to successfully feature in a sound explanation according to the D-N model, it would have to say only true things about the world. Much the same goes for causal accounts of explanation. To explain a token event or observation is to specify its causes; to explain a recurring scientific phenomenon is to cite the types of causes responsible for it, i.e. the underlying causal mechanisms. For a causal explanation to be successful, then, the cited causes must exist. As Nancy Cartwright puts it: "An explanation of an effect by a cause has an existential component, not just an optional extra ingredient." (Cartwright 1983, p. 91) Michael Strevens, specifically in relation to model-based explanations, claims that "no causal account of explanation […] allows non-veridical models to explain" (Strevens 2009, p. 320).

At the same time, it is a well-worn platitude that most models in science – not just those in economics – are, literally, false. Indeed, the very point of constructing models is to leave out an enormous amount of detail, and doing so ignores a great many entities that would otherwise have to be part of a full causal story. And it is not as though what remains is veridical either: by simplifying the relations and interactions between those entities and processes that *are* accounted for, using heavy-handed idealization and the like, models depart from reality in significant ways, even as they purport to represent a tiny slice of it. Models, more often than not, are not faithful representations of reality, but at best are heavily distorted caricatures of a limited segment of the world at large.

From the point of view of full-blown truth and causal veridicality, models thus appear to be deficient, and it would seem that models simply cannot meet the high demands associated with scientific explanation. That is, if we insist that our explanations be grounded in truth or reference to real existing causes, then we should not expect to get much explanatory mileage out of scientific models. Yet, from another angle, this seems to be exactly the wrong conclusion to draw. Models explain *in spite of* being highly idealized and limited in what they represent of reality. But even this does not seem quite right: after all, it is often not *in spite of*, but *because* models simplify, idealize, and generally highlight only a select few aspects of reality, that models do explain. More specifically, they are sometimes uniquely placed to shed light on real-world phenomena in ways that are not open to standard methods of explanation. This highlights the other side of the aforementioned tension in philosophical debates about models: the sense that models are not being given enough credit for the things they *can*, in fact, already achieve.

For example, as we shall see, models can be an excellent tool for constructing *how-possibly* explanations. The basic idea goes back to the work of William Dray in the 1950s and has recently been picked up by a number of contemporary philosophers of science. (See e.g. Forber 2010; Reydon 2012, and references therein.) Dray claimed to have identified a type of explanation that did not conform to the rationale behind the deductive-nomological account. Whereas D-N-type explanations demonstrate, with logical necessity, why an explanandum had to happen – because the explanandum follows logically from the (true) explanans – Dray resisted such, as he put it, "*why-necessarily*" explanations. Instead, he aimed for something more modest: namely the rebuttal of the "presumption that [the explanandum] could not have happened, by showing that, in the light of certain further facts, there is after all no good reason for supposing that it could not have happened" (Dray 1957, p. 161). Note the implicit double negative: Dray *rebuts* the thought that a certain explanandum *could not* have happened, so that, in succeeding with his rebuttal, he would have demonstrated that, for all we know, the explanandum had been possible all along.

Dray claimed that, when it comes to D-N type why-necessarily explanations and how-possibly explanations, the "two kinds of explanation are logically independent" (1957, p. 167) – though what he probably should have said is that they aim for different sorts of *explanatory insight*. Importantly, how-possibly explanations are not just – and are not intended to be – merely incomplete D-N type explanations. Rather, each type of explanation is a response to a distinct question, which in turn is emblematic of a distinct orientation of inquiry. In the case of why-necessarily explanations, the question we seek to answer is "Why is it so?" – the assumption being that circumstances, in conjunction with the relevant regularities and laws of nature, make the world thus and so. (This is why the preferred term, which is more suitable for use beyond the D-N model, is '*how-actually* explanations'.) By contrast, in the case of how-possibly explanations, we ask ourselves "How could it be?" – that is, we seek to identify a *possible* pathway of events or dependencies that, *if true*, could explain what we are observing.

A how-actually explanation can rest content once it has specified the circumstances and actual regularities that obtain and so has little need for explicitly including modal considerations. Within the D-N account, for example, the question of 'what could be' is

framed at best via the issue of prediction – yet, as is well-known, from the deductive-nomological standpoint predictions are structurally identical to explanations, so no genuinely new modal information is gained. To the extent that laws are being invoked, these may be taken to underwrite counterfactual inferences – though, of course, on Hempel's own view, laws are just regularities that meet certain additional requirements – a view that reflects a general "unwillingness to employ modal concepts as primitives" (Woodward 2009). How-possibly explanations, by contrast, place the modal dimension front row and centre, since they first invite us to consider a range of *ways the world might be*, before making the case that least one of the scenarios may, for all we know, obtain.

The twin questions of "Why is it so?" and "How could it be?" reflect different orientations toward the world-at-large. Of course, any successful answer to the former question also simultaneously answers the latter, since, *a fortiori*, any true account of how things are tells us how they are possible. However, one would seriously misunderstand the point of how-possibly explanations if one were to assimilate them to (just another kind of) potential explanations. As Dray rightly put it, how-possibly explanations amount to a successful *rebuttal* of the presumption that an event or phenomenon could not have happened. This takes on a special urgency when the existence of an event or phenomenon is disputed, perhaps by some argument to the effect that it could not possibly exist. In such a situation, constructing a how-possibly explanation may well serve the pragmatic goal of asserting the very existence of a disputed event or phenomenon – yet it does so by forcing us to consider a range of possible worlds and our relation to them.

3 Representational Ideals

Different orientations toward the world-at-large manifest themselves at various levels in science, from explanation and theory construction all the way to experimentation and model-building. Michael Weisberg has introduced the useful notion of *representational ideals*, which "regulate which factors are to be included in models, set up the standards theorists use to evaluate their models, and guide the direction of theoretical inquiry" (Weisberg 2007, p. 648). Such representational ideals have a primarily regulative function, in that they "do not describe a cognitive achievement that is literally possible", but instead "give the theorist guidance about what she should strive for and the proper direction for the advancement of her research program" (p. 649).

Some of the most familiar representational ideals are also the most general ones and correspond to widely acknowledged – though typically not simultaneously realizable – goals of science. Consider the ideal of COMPLETENESS, according to which "each property of the target phenomenon must be included in the model" (this is the *inclusion rule*) and, moreover, "with an arbitrarily high degree of precision and accuracy" (*the fidelity rule*; Weisberg 2007, p. 649). COMPLETENESS exhorts its adherents to "always add more detail, more complexity, and more precision" (ibid.), thereby functioning as a regulative ideal that can guide inquiry through various turns and twists. At the same time, it is clear that in actual scientific practice – given the obvious constraints in terms of time, access, and representational resources – completeness can never, ever be

achieved. Furthermore, the ideal of COMPLETENESS has to compete with other, *prima facie* equally plausible contenders, such as SIMPLICITY.

A less obvious representational ideal is what Weisberg calls 1-CAUSAL, which "instructs the theorist to include in the model only the core or primary causal factors that give rise to the phenomenon of interest" (Weisberg 2007, p. 654). Again, this is in obvious tension with COMPLETENESS. Nonetheless, as we shall see, it captures an important strand within the practice of scientific modelling. In many scientific contexts, we are indeed interested in representing only a small number of causal factors in a given (actual) target system. 1-CAUSAL, in such a situation, recommends that we should leave out factors of secondary importance, focusing instead only on the core causal features of a system. Note, however, that 1-CAUSAL can also be put to other uses. Instead of looking at actual target systems and simplifying them by leaving out non-core causal factors, we can also turn our attention to causal mechanisms of interest and consider what a system (or world) *would look like*, if it were governed only (or primarily) by the mechanisms in question. In other words, we can look towards 1-CAUSAL not as a regulative ideal, but as a guide for constructing new model-worlds.

That different regulative ideals, such as COMPLETENESS and SIMPLICITY, trade off against each other, and for all sorts of reasons, is nicely illustrated by a recent exchange between Elizabeth Lloyd and Wendy Parker, concerning what should guide the development and interpretation of models in climate science. On the issue of confirmation, Lloyd argues that instances of fit between the output of climate models and observational data confirm the model as a whole, thereby boosting its overall empirical adequacy. Though Lloyd does not say so explicitly, one natural move is to infer a commitment on her part to the idea that the more complete a climate model is, the better – in the long run – its fit with empirical data is going to be. Wendy Parker criticizes just that and argues that we must shift from a focus on empirical adequacy to "the adequacy of climate models for particular purposes", that is *adequacy-for-purpose*. On this account, a better model might be one that is better at predicting important events such as droughts – even if it gets many more of the less-interesting aspects wrong.

Different representational ideals prize different aspects of an explanation, theory, or model. In many cases, as in the example of COMPLETENESS and SIMPLICITY, the corresponding desiderata cannot simultaneously be maximized, but instead trade off against each other. Representational ideals also differ in the way that modal considerations enter into the picture. Take generality which, other things being equal, is a desideratum of most models. As Weisberg has argued, GENERALITY in the context of models refers to at least two different characteristics, *A-generality* and *P-generality*: "A-generality is the number of actual targets a particular model applies to given the theorist's adopted fidelity criteria. P-generality, however, is the number of possible, but not necessarily actual, targets a particular model captures." (Weisberg 2007, p. 653) Whereas A-generality can be materially satisfied by a fluke, thanks to contingent features of the actual world, P-generality requires that we consider a (properly restricted) range of possible worlds – perhaps a subset of nomically possible worlds that are compatible with our background knowledge. When interpreted as a representational ideal, then, P-GENERALITY would seem to exhort us to include modal considerations in our assessment of scientific models, theories, and explanations.

4 Galilean and Minimalist Idealization

Model construction, where it aims to represent real target systems, is widely assumed to be a matter of *simplification*: irrelevant detail gets left out, and that which remains is subject to idealization. Sometimes, specific moves in the process of model construction – dropping a higher-order term or restricting interactions to a subset of elements, to mention but two recurring operations that can be routinely found in science – are being described as the 'introduction of an idealization', following which the model is said to 'contain' an idealization:

> "A causal model contains an idealization when it correctly describes some of the causal factors at work, but falsely assumes that other factors that affect the outcome are absent." (Elgin and Sober 2002, p. 448)

Recently, the idea that idealization applies only to specific 'ingredients' of a model, which may then be judged as to whether they are "harmless" (ibid.) or not, has come under fire. Idealization, Rice (2019) has argued, leads to "holistic distortion" of the model as a whole. For the purposes of this paper, I shall not take a stand on this issue, but shall instead focus on two of the most prominent strategies of idealization, *Galilean* idealization and *minimalist* idealization.

McMullin (1985) has provided an account of Galilean idealization, according to which distortions are routinely being introduced with the goal of simplifying theories and making the resulting models computationally tractable. For such idealized models to meet the stringent demands of scientific realism, there must be the realistic hope that future corrections, for example by adding detail back in, can show the distortions to be irrelevant. Indeed, this is where such advances as improvements in computational power come into play: with new techniques, the need for Galilean idealization will diminish, since the main justification for its use is pragmatic in the first place. Galilean idealization, thus understood, is a largely auxiliary 'stop-gap' measure, the goal of which is the later re-introduction of complexity; it is idealization *with the goal of de-idealization*. As Weisberg puts it, "Galilean idealization takes place with the expectation of future deidealization and more accurate representation" (Weisberg 2007, p. 642).

This characterization, which I take to be a fair one, already contains an important clue regarding the role of modal considerations in Galilean idealization. Since de-idealization is supposed to re-create (or at least approximate) the richness of the *actual* empirical target phenomenon, for Galilean idealization to be compelling, it can never venture all that far from the actual world. Radically counterfactual scenarios, for example, would not be amenable to Galilean idealization, since there would be no determinate sense in which the adding of detail could ever lead to a 'de-idealization' of the model. In fact, the downplaying of modal considerations is evident from Galileo's own pronouncements, when he argues that, since we cannot "investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistances", we should instead "observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant" (quoted after Weisberg 2007, p. 641). The measure of the success of Galilean idealization is the

extent to which it is able to be linked back again to those systems that are, in fact, accessible from within the actual world.

The situation is markedly different for *minimalist idealization*, which is Weisberg's term for labelling "the practice of constructing and studying theoretical models that include only the core causal factors which give rise to a phenomenon" – the core causal features being those (and only those) that "*make a difference* to the occurrence and essential character of the phenomenon in question" (Weisberg 2007, p. 648). It would not be inaccurate, in the light of our discussion in the previous section, to think of minimalist idealization as a methodological tool – perhaps *the* methodological tool – for implementing the representational ideal of 1-CAUSAL. In the next section, we will encounter three types of models – toy models, minimal models, and exploratory models – which rely heavily (though not in all cases constitutively) on minimalist idealization as a tool for model-building. It is worth emphasizing, however, that, as in any implementation of a representational ideal, minimalist idealization must happen against the backdrop of a (perhaps tacitly) shared understanding of what it is that we set out to model. In particular, what is required is a prior judgment as to which elements properly belong to the core of a system, or which aspects of a phenomenon make up its "essential character". Such judgements will likely vary across disciplines, and failure to revise them when necessary – for example when adapting a model from one discipline to another – may well introduce a further, hidden source of distortion.

5 Minimal Models, Toy Models, Exploratory Models

With this caveat in mind, we can now turn to three types of models that have garnered growing philosophical attention in recent years: minimal models, toy models, and exploratory models. It should be emphasized that all three labels emerged independently and were in part co-opted from prior scientific usage, both of which introduces a certain degree of fuzziness into the concepts. Thus, by juxtaposing the three, I am not intending to offer a neat three-fold taxonomy, nor do I wish to endorse or defend any conceptual redundancy or proliferation of terms that may result from this juxtaposition. Rather, the present discussion is motivated by the thought that all three labels latch on to important features, and recurring strategies, of the practice of scientific modelling.

5.1 Minimal Models

The first type of models to be discussed is the case of *minimal models*, which has received some philosophical prominence via Robert Batterman's work on asymptotic reasoning, especially in statistical physics. Interestingly, Batterman draws inspiration from a quote by a practicing physicist, Nigel Goldenfeld, who uses the phrase "the correct minimal model" to refer to "that model which most economically caricatures the essential physics" (Goldenfeld 1992, p. 33). Elsewhere in science, similar terminology has been developed, for example by theoretical ecologists, who occasionally speak of "a minimal model for ideas", by which they mean a model that "is intended to explore a concept without reference to a particular species or place" (Roughgarden et al. 1996, p. 26). As this usage suggests, minimal models are *not* merely cleaned-up

models of actual target systems, developed with an eye towards future de-idealization, but are being investigated in their own right, as models that shine a spotlight on what we earlier called the "essential character" of a phenomenon, or class of phenomena.

Batterman, in connection with minimal models in statistical physics, speaks of "*highly idealized minimal models* of the universal, repeatable feature of a system" (Batterman 2002, p. 36) The term 'universal' here adverts to the *universality classes* into which different physical systems – indeed, very different systems, judging only by their object-level features! – can be grouped, based on whether they share the same scale-invariant limit under the process of renormalization group flow. The general idea is perhaps easiest to grasp by way of example. Consider a system consisting of many interacting entities, such as atoms and electrons in a crystal lattice. Such a system may exhibit very salient macroscopic behaviour, such as transitioning to an orderly (e.g., magnetic) state below a certain transition temperature. The outwardly observable macroscopic behaviour is the collective outcome of the interactions between the micro-constituents of the system, yet due to their large number it would be a hopeless task to try to calculate the combined effect of all the individual constituents directly. A different, more promising strategy, pioneered by Leo Kadanoff (1966), would attempt to 'bundle up' individual components (for example groups of nearest neighbours on a lattice), 'average out' their dynamic behaviour, and map it – through a process of 're-scaling' – onto a single element in a higher-level representation of the system. (See Fig. 1.) This radically reduces the number of individual elements one has to consider in the new system, and since the process can be repeated again for the new system, each iteration will render the resulting system more tractable.

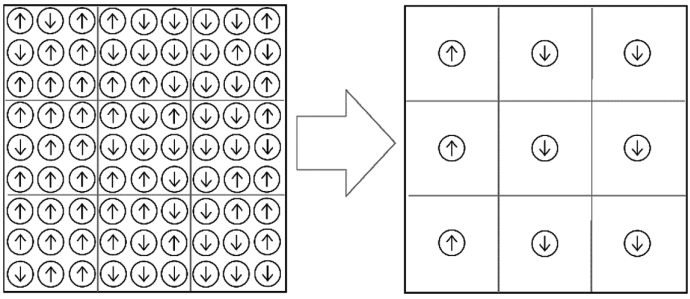


Fig. 1. Example of a 'block spin transformation' mapping a (3 × 3) group of components in the original system on the left onto a single element in the higher-level representation on the right.

While details of the target system will inevitably be lost in this process of 'renormalization', many phenomena – such as the critical behaviour near a phase transition – are known to depend less on individual variations at the constituent level and more on overall features of the system's structure and the type of interaction between its elements. For many purposes, the loss of detail thus turns out to be a blessing in disguise: basically, degrees of freedom that are irrelevant are being systematically eliminated, leaving behind only a small number of characteristics that

nonetheless appear to govern, amongst others, the thermodynamic behaviour under phase transitions (e.g. from solid to liquid, etc.). As Batterman and Rice put it:

> The idea is to construct a space of possible systems. This is an abstract space in which each point might represent a real fluid, a possible fluid, a solid, and so on. […] By rescaling, one takes the original system to a new (possibly nonactual) system/model in the space of systems that exhibits continuum scale behavior similar to the system one started with. (Batterman and Rice 2014, p. 362)

A minimal model, in Batterman's sense, is not – and is not intended to be – merely an 'impoverished' representation of an actual target system. On the contrary, insofar as it represents its target *as a member of a universality class*, it does so in an admirably concise manner. For the purposes of representing its critical behaviour near a phase transition, say, there may simply be nothing more that we need to know. The very idea – so crucial to Galilean idealization – that we can re-create the empirical richness of the target system through *de-idealization* does not apply to the practice of deploying minimal models, since "the adding of details with the goal of 'improving' the minimal model is self-defeating" (Batterman 2002, p. 26).

Once again, acknowledging the modal dimension of modelling helps to make sense of why minimal models – in the right sorts of contexts – are so useful. If we wish to understand why *this* particular target system displays the critical behaviour it does, then realizing that it belongs to a universality class of systems that all share the same behaviour in the vicinity of a phase transition, goes some way towards assuring us that, even allowing for minor variations due to our imperfect background knowledge, we could not easily be wrong about the qualitative behaviour of the target. Even if we can only ever have incomplete knowledge of any given target system, we may still know just enough to be able to derive, and explain, those aspects of the target system that interest us. Indeed, as Batterman and Rice note, "[t]he renormalization group strategy, in delimiting the universality class, provides the relevant modal structure that makes the model explanatory: we can employ complete caricatures—minimal models that look nothing like the actual systems—in explanatory contexts because we have been able to demonstrate that these caricatures are in the relevant universality class." (Batterman and Rice 2014, p. 264)

5.2 Toy Models

As in the case of minimal models, the practice of using toy models offers yet another perspective on model-building beyond the narrow goal of representing a specific actual target system. Scientists frequently invoke the creative freedom that such modelling approaches afford them. In her work on scientists' thoughts on scientific models, Bailer-Jones (2002) quotes from an interview with a theoretical solid-state physicist, John Bolton, who expresses a widely held sentiment among modellers, when he argues that the model world is "not the real world": "It's a toy world, but you hope it captures some central aspects of reality, qualitatively at least, how systems respond" (cited after Bailer-Jones 2002, p. 295). A toy model, on this account, amounts to the creation – by fiat – of a model-world that operates according to stipulated rules, which may well be motivated by interest in a real target system, but whose validity is not – certainly not

initially – to be judged by whether it constitutes an empirically adequate representation of that target system. As has sometimes been noted, if a toy model, in this sense, "cannot be regarded as a model of a (real) target system", then "its epistemic value seems doubtful" (Gottschalk-Mazouz 2012, p. 17); it is only by dropping the assumption that a model must always be "depicting reality" that the role of toy models within the practice of scientific modelling becomes intelligible.

At one extreme, it has been argued that, categorically, "toy models do not perform a representational function" (Luczak 2016, p. 1). According to this view, it is part of the definition of a toy model that it must lack a specific target (where this may be either a particular system or, as in the case of minimal models, a class of physical systems). Instead, toy models are thought to mainly serve as a training ground for "certain formal techniques" or for "elucidat[ing] certain ideas relevant to a theory" (ibid.).[1] This, however, seems too strong, for there does not appear to be any principled reason why, on occasion, toy models should not succeed in representing a real target system, even if this is not their main intended function.

Alexander Reutlinger, Dominik Hangleiter, and Stephan Hartmann, in a 2018 paper, make a useful distinction between "embedded" and "autonomous" toy models. Whereas embedded models are "models of an empirically well-confirmed framework theory" (Reutlinger et al. 2018, p. 1072), autonomous models are not derived from (and perhaps cannot be derived from) an underlying theory. As an example of the former, consider modelling the orbit of a single planet – say, Earth – around the Sun using the equations of Newtonian mechanics. As a representation of the solar system, such a model would be an empirical failure (since, by definition, it neglects all the other celestial bodies, including the Moon and the planets), and even on theoretical grounds, one might object that the model is based on an outdated theory, Newtonian physics having been superseded by Einstein's relativity theory. Yet it is clear that, for many purposes – practical and theoretical alike – one stands to learn quite a bit from considering such a highly simplistic, strongly idealized system. In particular, one can learn a lot from it about, say, gravitational objects on a closed orbit according to Newtonian mechanics. Its formal character as an instantiation of Newtonian mechanics is what makes such a model an *embedded* toy model. By contrast, Schelling's model of segregation, which is often interpreted as a model of racial segregation in urban areas, operates outside any established framework theory. Instead it makes various basic assumptions – two types of agents (black and white) distributed randomly on a grid, following clear behavioural rules (e.g., 'randomly move to another location if less than a third of your neighbours are of the same colour as you') – which are neither deduced from an underlying theory nor inferred from data, but instead are simply *posited* as true in the 'toy world' of the model. Its independence from an underlying theory renders the Schelling model an *autonomous* toy model.[2]

[1] Luczak also notes the heuristic function of "generat[ing] hypotheses about other systems", which, on my interpretation, would best be subsumed under *exploratory uses* of models, to be discussed in Sect. 5.3 of this paper (see also Gelfert 2016, ch. 4).

[2] See (Reutlinger, Hangleiter, and Hartmann 2018, pp. 1075–1077).

What toy models have in common, according to Reutlinger, Hangleiter, and Hartmann, is the following three features:

1. They "are strongly idealized in that they often include both Aristotelian and Galilean idealizations";
2. they "are extremely simple in that they represent a small number of causal factors […] responsible for the target phenomenon"; and
3. they "refer to a target phenomenon". (Reutlinger et al. 2018, p. 1070)

That is, contra Luczak (2016), they explicitly build representational success into their definition of the term 'toy model'. Again, however, it seems to me that there is no principled reason why one must take a stance on this issue. One could speculate whether the subjective need to do so, which is reflected by Luczak's dismissal, and Reutlinger, Hangleiter, and Hartmann's endorsement, of a representational function of toy models, echoes an influential tradition within philosophy of science, which takes representation to be prior to any consideration of the methods and tools of actual science. Perhaps, then, it is time to free models not only from their role as intermediaries between theory and data (and begin to regard them as standing "outside the theory–world axis"; Morrison and Morgan 1999, p. 18), but also from the primacy of representation. Some toy models may represent actual target systems, others do not – and yet, they are not therefore useless to science (for the very reasons discussed above).

Before turning to the case of exploratory models in more detail, it is worth noting some overlap with the discussion of toy models so far. *Autonomous* toy models are of particular interest in this regard, since they do not require – and indeed are taken to be autonomous from – any underlying theory that may or may not exist concerning the phenomenon in question. As we shall see, exploratory models have their paradigm domain of application in contexts where we lack a fully-formed (or readily available) underlying theory. In such a situation, modelling may serve the purpose of developing a grasp of an (as yet theoretically inaccessible) phenomenon, or it may even lead us to reconsider whether we are dealing with a unified and coherent (that is, empirically stable) phenomenon in the first place. Finally, there exists more than a passing resemblance between toy models as discussed so far and what Robert Sugden, in connection with economic models, has called "credible world" modelling. Rather than begin with our best description of an actual system and then gradually derive simplified models from it, Sugden argues we can often *stipulate* how a model world ought to behave: "The model world is not constructed by starting with the real world and stripping out complicating factors: although the model world is simpler than the real world, the one is not a *simplification* of the other." (Sugden 2000, p. 25) That is, there is a clear constructive – indeed, *imaginative* – element to model-building. Testing such a model's credibility first requires ascertaining whether it coheres internally. Whether it does or not is a matter of whether its results "can be seen to follow naturally from a clear conception of how the world might be" (p. 26). Achieving clarity about *possible* relationships in the model world thus is prior to comparing the model against empirical reality. Only in a second step does the model's relation to the actual world need to be considered: "For a model to have credibility, it is not enough that its assumptions cohere with one another; they must also cohere with what is known about causal processes in the real world" (ibid.).

5.3 Exploratory Models

The notion of 'exploratory modelling', as used in philosophy of science, was initially motivated by analogy with 'exploratory experimentation', which has received considerable attention, especially from historians of science since the mid-1990s. The introduction of the latter was itself a reaction against a narrow view that dominated philosophical discussions concerning the relation between theory and experiment. (See Gelfert 2016, ch. 4.) According to this view, experiments primarily serve to test scientific theories and hypotheses. That is, "in those cases that [had] traditionally received the most attention in philosophy of science, a significant prior body of theoretical knowledge [was] assumed to be available" (Gelfert 2018, p. 258). Yet, often enough – especially during the early phases of scientific inquiry – the existence of an integrated body of theoretical knowledge cannot be assumed, either because such knowledge is not readily available or because it is itself a matter of dispute. The label 'exploratory' is meant to capture just such episodes of scientific inquiry; that is, situations where the stability of the putative target phenomenon has not yet been ascertained in ways that lend themselves to theoretical description using shared and accepted principles and concepts. While such situations will occur most obviously during the initial stages of research, it is important to think of the term not merely as a temporal marker. For one, theoretical indeterminacy can last well beyond the initial steps of research; moreover, the very subject matter of interest may not allow for an underlying 'fundamental theory'.

As a case in point consider early models of traffic flow in 20th-century sociodynamics. Early such models looked toward fluid dynamics for inspiration, yet, perhaps unsurprisingly, were not successful at capturing various central features of vehicular traffic flow. By the 1950s, it had become clear that any successful model of vehicular traffic flow would need to account for a variety of quite heterogeneous factors, ranging from physical quantities such as acceleration, speed, and length of the vehicles all the way to psychological factors such as the drivers' reaction time. It was not until 1953 that the American engineer Louis Pipes (1953) proposed the first car-following model, which modelled car traffic as the cumulative effect of each driver responding to the car in front of her.[3] It is clear that, simply in virtue of subject matter, Pipes could not draw on an underlying fundamental theory – since any such theory would have to integrate wildly disparate phenomena. There simply is not, and presumably never will be, a 'fundamental' theory that accounts equally for the speed of a car and for its driver's reaction time. Yet the ability of Pipes' model to account, at least in principle, for the spontaneous formation of traffic jams led to a proliferation of subsequent car-following models, and the bold step of positing an exploratory model more than paid off, since it provided a fruitful starting point for future quantitative study of the complex phenomenon of car traffic.

Yet exploratory models do not merely serve the heuristic function of stimulating subsequent research. Often enough, they explicitly aim at identifying how-possibly explanations or otherwise delineate the space of possibilities. This is nicely illustrated

[3] For a discussion of this example as an illustration of one of several key functions of exploratory modelling, see (Gelfert 2016, pp. 85–86).

by an example from theoretical biology, Alan Turing's – then entirely speculative – proposal of reaction-diffusion processes as a causal basis of biological pattern formation.[4] The basic idea is that cell differentiation in biological systems, and the subsequent development of spatially distinct structures in an organism, may be the result of the interplay between two 'morphogens', i.e. two biochemically produced substances that diffuse at different speeds, one of which is locally activated, whereas the other gives rise to long-range inhibition. As a result of the different diffusion rates, Turing's model predicts varying concentrations of the two morphogens according to a 'chemical wavelength', depending on the organism's boundary conditions, which in turn may trigger the expression of different phenotypes. Turing was careful to stress that he did not wish to "make any new hypotheses" of a biologically substantive kind, but only wanted to make the case for "a possible mechanism by which the genes of a zygote may determine the anatomical structure of the resulting organism" (Turing 1952, p. 37). For Turing, identifying a *possible* mechanism sufficed to show "that certain well-known physical laws are sufficient to account for many of the facts" (ibid.) of biological form, thereby proving a point concerning the fundamental character of biological phenomena rather than representing any empirical target system in particular.

As these examples already suggest, it would be wrong to think of exploratory models as a unified class in virtue of some intrinsic features of the model. Like autonomous toy models, they are models that are not (and perhaps cannot) be embedded into a well-confirmed underlying empirical theory; unlike minimal models, they do not bear any special affinity to specific types of (e.g. asymptotic) reasoning in particular subdisciplines. This suggests that a proper consideration of exploratory models offers a complementary perspective to the one afforded by looking at minimal models and toy models, respectively. In particular, juxtaposing all three types of models highlights the continuity that exists between the exploratory stages and what one might call the 'mature phase' of scientific research. Many of the central features of toy models and minimal models – a concern for identifying core mechanisms and processes (rather than aiming for full empirical adequacy), a stipulative element that goes well beyond traditional notion of 'idealizing a target system', and a liberal interpretation of the goal of scientific representation (which is not limited to individual target systems, but extends to classes of systems as well as counterfactual scenarios) – have a natural justification in exploratory contexts.

Furthermore, one finds an equal – or at least comparable – concern with *modal considerations* across all three types of models. Thus, one of the core functions of exploratory models is that of providing a "proof of principle" (e.g. in the form of a how-possibly explanation). At the same time, exploratory models delineate the space of what is possible by also exploring *impossibilities*. Consider the counterfactual case of three-sex species, which can be shown to be evolutionarily unstable using spatial (celullar automata) models of mating: "[T]hree-sex species are not idealization[s] about anything real, but fictions not accessible by means of simplification or abstraction

[4] For a full discussion of this example from the perspective of exploratory modelling, see (Gelfert 2018).

carried out on real systems." (Diéguez 2015, p. 171) A similar point regarding the importance of delineating possibilities and impossibilities is advanced by Michela Massimi in her discussion of the exploratory role of what she calls 'perspectival models'. In such situations, she argues, the representational content of models "is not about mapping onto an actual worldly-state-of-affairs (or suitable parts thereof) but has instead a *modal* aspect: it is about exploring and ruling out the space of possibilities in domains that are still very much open-ended for scientific discovery." (Massimi 2018, p. 338) Whereas Massimi aims for a reconciliation between science's plurality of practices, of which exploratory modelling is just one example, and the goals of scientific realism, the goal of the present section was a more modest one: to vindicate exploration, and in particular its modal dimension, as one of the core functions of scientific modelling, on a par with the familiar goals of representational success, empirical adequacy, prediction and explanation.

6 Conclusion

I have argued that all three types of models discussed in this paper – minimal models, toy models, and exploratory models – are well-suited to the study of modal characteristics across a wide range of phenomena, target systems, and (actual or prospective) theories. This, in itself, may be a weak claim, but it suggests further lines of inquiry. For example, it seems plausible to think that models of the three types discussed may well outperform data-driven models that are based purely on fit with past measurements and data. Likewise, it would be an interesting project to track the various uses of toy models and minimal models across different stages in the development of specific scientific debates. Given the convergences that exist between minimal models, toy models, and exploratory models, it seems plausible to think that toy models – and *autonomous* toy models, in particular – should be among the preferred types of models during those phases of scientific inquiry that are characterized by the absence of a fully-formed and widely accepted 'underlying theory'.

These are empirical hypotheses, which only a detailed analysis based on a range of case studies from the history of science would be in a position to answer. Here, I wish to restrict myself to a (perhaps speculative) sketch of an argument in favour of the *prima facie* plausibility of such hypotheses. During mature phases of science, when well-confirmed theories are readily available, gaining modal knowledge is tantamount to acquiring a deeper understanding of why things are thus and so. As an example, consider the case of phase transitions as described by statistical physics. Against the backdrop of such a well-established theoretical account, our ability to subsume various (actual and possible) systems under universality classes, along with our figuring out why certain minimal models can reproduce such thermodynamic behaviour, deepens our understanding of actual systems. By being able to locate the actual systems we are studying in the space of possibilities, we gain knowledge about what it would take for things to be different (and why, given the circumstances, no other empirical findings were to be expected). By contrast, during periods of inquiry that are dominated by *exploratory* concerns, we will often be uncertain as to *which* scenario, among a range of possible worlds compatible with our limited background knowledge, we find

ourselves in. In such a situation, it is eminently rational to use models in order to probe the space of possibilities – not so much in order to deepen our understanding of what we already know, but rather in order to figure out what type of world we are actually in.

To be sure, the two cases are not entirely symmetrical and are subject to a host of competing epistemic and non-epistemic interests and constraints. Yet, I believe they both indicate a need to gain – and exploit – modal information, a need that can often be satisfied using the types of models discussed in this paper. While this is no more than a sketch of an argument, it may suffice as an illustration of the general observation that models are as much about *exploring what there could be* as they are about *representing what there is*.

References

Bailer-Jones D (2002) Scientists' thoughts on scientific models. Perspect Sci 10(3):275–301
Batterman R (2002) Asymptotics and the role of minimal models. Br J Philos Sci 53(1):21–38
Batterman R, Rice C (2014) Minimal model explanations. Philos Sci 81(3):349–376
Cartwright N (1983) How the laws of physics lie. Oxford University Press, Oxford
Diéguez A (2015) Scientific understanding and the explanatory use of false models. In: Bertolaso M (ed) The future of scientific practice: 'bio-techno-logos'. Pickering & Chatto, London, pp 161–178
Dray W (1957) Laws and explanation in history. Clarendon Press, Oxford
Forber P (2010) Confirmation and explaining how possible. Stud Hist Philos Biol Biomed Sci 41 (1):32–40
Gelfert A (2016) How to do science with models: a philosophical primer. Springer, Cham
Gelfert A (2018) Models in search of targets: exploratory modelling and the case of Turing patterns. In: Christian A, Hommen D, Retzlaff N, Schurz G (eds) Philosophy of science: between natural sciences, social sciences, and humanities. Springer, Dordrecht, pp 245–271
Goldenfeld N (1992) Lectures on phase transitions and the renormalization group. Addison-Wesley, Boston
Gottschalk-Mazouz N (2012) Toy Modeling: Warum gibt es (immer noch) sehr einfache Modelle in den empirischen Wissenschaften? In: Fischer P, Luckner A, Ramming U (eds) Die Reflexion des Möglichen. LIT-Verlag, Berlin, pp 17–30
Kadanoff LP (1966) Scaling laws for Ising models near T_c. Physics 2(6):263–272
Luczak J (2016) Talk about toy models. Stud. Hist. Philos. Mod. Phys. 57(1):1–7
Massimi M (2018) Perspectival modeling. Philos Sci 85(3):335–359
McMullin E (1985) Galilean idealization. Stud History Philos Sci Part A 16(3):247–273
Mehmet M, Sober E (2002) Cartwright on explanation and idealization. In: Earman J, Glymour C, Mitchell S (eds) Ceteris paribus laws. Kluwer, Dordrecht, pp 165–174
Morrison M, Morgan M (1999) Models as mediating instruments. In: Morrison M, Morgan M (eds) Models as mediators: perspectives on natural and social science. Cambridge University Press, Cambridge, pp 10–37
Pipes LA (1953) An operational analysis of traffic dynamics. J Appl Phys 24(3):274–281
Reiss J (2012) The explanation paradox. J Econ Methodol 19(1):43–62
Reutlinger A, Hangleiter D, Hartmann S (2018) Understanding (with) toy models. Br. J. Philos. Sci. 69(4):1069–1099
Reydon T (2012) How-possibly explanations as genuine explanations and helpful heuristics: a comment on Forber. Stud Hist Philos Biol Biomed Sci 43(1):302–310

Rice C (2019) Models don't decompose that way: a holistic view of idealized models. Br J Philos Sci 70(1):179–208

Roughgarden J, Bergman A, Shafir S, Taylor C (1996) Adaptive computation in ecology and evolution: a guide for future research. In: Belew RK, Mitchell M (eds) Adaptive individuals in evolving populations: models and algorithms. Addison-Wesley, Boston, pp 25–30

Strevens M (2009) Depth: an account of scientific explanation. Harvard University Press, Cambridge

Sugden R (2000) Credible worlds: the status of theoretical models in economics. J Econ Methodol 7(1):1–31

Turing A (1952) The chemical basis of morphogenesis. Philos Trans Roy Soc Lond (Ser B Biol Sci) 237(641):37–72

Weisberg M (2007) Three kinds of idealization. J Philos 104(12):639–659

Woodward J (2009) Scientific explanation In: Zalta E (ed) Stanford Encyclopedia of Philosophy (Spring 2009 Edition). https://stanford.library.sydney.edu.au/archives/spr2009/entries/scientific-explanation/. Accessed 03 Mar 2019



Model Types and Explanatory Styles in Cognitive Theories

Simone Pinna(✉) and Marco Giunti

Dipartimento di Pedagogia, Psicologia, Filosofia, ALOPHIS (Applied LOgic, Philosophy and HIstory of Science), Università degli Studi di Cagliari, Via Is Mirrionis, 1, Cagliari, Italy
simonepinna@hotmail.it, giunti@unica.it

Abstract. In this paper we argue that the debate between representational and anti-representational cognitive theories cannot be reduced to a difference between the types of model respectively employed. We show that, on the one side, models standardly used in representational theories, such as computational ones, can be analyzed in the context of dynamical systems theory and, on the other, non-representational theories such as Gibson's ecological psychology can be formalized with the use of computational models. Given these considerations, we propose that the true conceptual difference between representational and anti-representational cognitive descriptions should be characterized in terms of *style of explanation*, which indicates the particular stance taken by a theory with respect to its explanatory target.

Keywords: Cognitive explanations · Computationalism · Dynamical approach · Ecological psychology

Introduction

The contrast between representational and non-representational theories in psychology and cognitive science is often associated with intrinsic differences between the types of model respectively employed. In this sense, the choice of a specific type of model for the description of a certain cognitive phenomenon (or a set of phenomena) will directly influence the kind of explanation, representational or non-representational, that is given to that phenomenon or set of phenomena by the theory. For example, the difference between computationalism and the dynamical approach to cognition has been read by many scholars precisely in these terms. In this view, the rejection of purely representational explanations, like those provided by computational models of cognition, is only possible by adopting a totally different kind of model, whose interwoven parts cannot be said to internally represent any external datum in the traditional sense.

In this paper, we argue that there is no direct connection between the kind of cognitive explanation given by a theory and the type of model employed by

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 20–40, 2019.
https://doi.org/10.1007/978-3-030-32722-4_2

the theory for the description of some cognitive phenomenon. To show this, we describe examples of cognitive theories where the connection above is not present. This means that the source of the difference between representational and non-representational cognitive explanations is to be found elsewhere. To this end, we introduce the notion of *style* of explanation, through which it is possible to understand the real connection between model type and kind of explanation.

1 Cognitive Systems as Dynamical Systems

Some presuppositions of classic computational cognitive science, in particular the central role attributed to inner representations for the explanation of cognitive processes, have been challenged by the so-called dynamical approach to cognition. In van Gelder and Port (1995) this approach is briefly summarized as a *dynamical hypothesis*: "natural cognitive systems are dynamical systems" (van Gelder and Port 1995, p. 11). What are, then, dynamical systems? To answer this question, it is necessary to define what a system is. The same authors proposed the following informal definition:

> [A] system is a set of changing aspects of the world. The overall state of the system at a given time is just the way these aspects happen to be at that time. The behavior of the system is the change over time in its overall state. [...] Not just any set of aspects of the world constitute a system. A system is distinguished by the fact that its aspects somehow belongs together. This really has two sides. First, the aspects must interact with each other; the way any one of them changes must depend on the way the others are. Second, if there is some further aspect of the world that interacts in this sense with anything in the set, then clearly it too is really part of the same system (van Gelder and Port 1995, p. 5).

A dynamical system is a special kind of system in which the interdependence between different parts, i.e. between its components, is expressed by some law of behavior. The overall state of the system at a given time (instantaneous state) is characterized by the value of its components at that time, and the set of all possible states of the system constitutes its state space – or phase space. Van Gelder provides a sufficiently broad informal definition of a dynamical system:

> A dynamical system for current purposes is a set of quantitative variables changing continually, concurrently, and interdependently over quantitative time in accordance with dynamical laws described by some sets of equations (van Gelder 1999, p. 245).

In van Gelder's view, the opposition between dynamical and computational models of cognition is irreconcilable, for the two approaches presuppose radically different concepts of time (the real/continuous time of actual physical systems vs the discrete time of computational steps). Before discussing this rather problematic point, it is useful to describe the chief example used in van Gelder (1995) to show the difference between the two approaches.

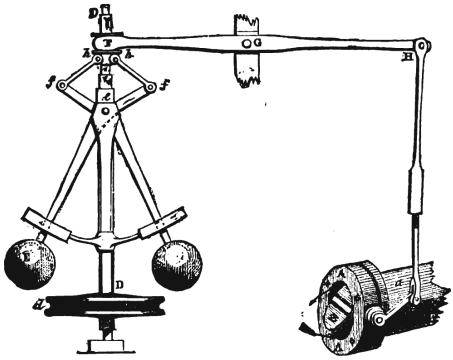


Fig. 1. Sketch of a Watt's centrifugal governor linked to the throttle valve (Image from "Discoveries & Inventions of the Nineteenth Century" by R. Routledge, 13th edition, 1900.)

1.1 The Governor's Problem

After the invention and the first improvements of the Watt steam engine, the steam power could finally be applied to any flywheel driven machinery. The main problem to solve, then, was finding a way to keep the turning speed of the flywheel constant. This could be done by real-time adjustment of the throttle valve that regulates the flux of steam from the boiler to the piston. However, the permanent employment of a human mechanic to do this work was unprofitable and, moreover, manual adjustment may not be sufficiently precise. So the problem was to design a device (governor) to do this work. The problem this governor had to solve may be algorithmically sketched as follows:

1. Measure the speed of the flyweel.
2. Compare the actual speed against the desired speed.
3. If there is no discrepancy, return to step 1. Otherwise,
   a. measure the current steam pressure;
   b. calculate the desired alteration in steam pressure;
   c. calculate the necessary throttle valve adjustment.
4. Make the throttle valve adjustment.
   Return to step 1 (van Gelder 1995, p. 348).

One may think that the governor's problem may be tackled by using devices such as tachometers, pressure meters, and any kind of measuring tools and effectors needed to carry out all the algorithmic steps seen above. This could have been a possible effective solution to the problem, but presupposed the existence of quite complex computational devices, something that was far beyond the possibilities of the eighteenth century technology. The actual solution, taken from existing treadmill technology, was much more efficient and elegant. It consisted of a shaft rotating concurrently with the main flywheel. Attached to the shaft there were two arms, on the end of which was a metal ball. As the rotation speed of the shaft increased, the centrifugal force drove the balls outward and hence

upward. The arms were linked directly to the throttle valve, so that increase and decrease of rotation speed could be directly used to regulate the flux of steam (see Fig. 1).

According to van Gelder, the difference between the (possible) algorithmic description of the centrifugal governor and actual Watt's solution makes clear the general contrast between computational and dynamic explanations of cognitive phenomena. In particular, this example highlights the completely different weight given to the explanatory role of representations by the two approaches. In computational explanations, cognitive processes are viewed as algorithmic transformations of mental symbols that represent various kind of data (perceptual, proprioceptive data, and so on). In dynamical explanations, by contrast, representations do not have an explanatory role. We may say, for example, that the angle assumed by the arms in Watt's device "represents" the rotation speed of the flywheel, the same speed "represents" in turn the amount of steam flowing through the throttle valve, etc. But in these utterances the concept of representation assumes a mere metaphorical role, which is very different from the foundational one that this concept has in computational cognitive science. In the dynamical explanation of the functioning of Watt's governor, indeed, we can totally get rid of the concept of representation and describe the system just by specifying the dynamical law that connects its main components, namely, the speed of the flywheel (determining the angle assumed by the shaft arms) and the degree of the throttle valve opening. Van Gelder, assuming that cognitive systems are dynamical systems, proposes that dynamical explanations of this kind should be given of cognitive phenomena, too.

1.2 Watt's Governor and Styles of Explanation

As we have already pointed out, van Gelder's example is aimed to show the main differences between representational (computational) and non-representational (dynamical) *cognitive* explanations. Ideed, even if the description of the governor's functioning is not, *per se*, a cognitive one, it can be considered as an abstract example of a functional/mechanistic explanation, as any significant cognitive explanation is (or, at least, should be). In our view, this example gives us all the elements we need to individuate the different *styles of explanation* of representational vs non-representational theories, namely, the different stances taken by the respective theories with regard to their explanatory targets. The main features of these styles are specified by the following two points.

– *Causal vs metaphoric role of representations*: in the computational solution proposed for the governor's problem, representations have a precise causal role, in the sense that the system's trajectory through its computational steps is driven deterministically by the information taken from the various measuring devices monitoring the system. On the other hand, as we have seen, in the dynamical solution representations may be only used as metaphors to describe the system's behavior, having no direct causal role, because no part of the system is expressly designed to measure physical magnitudes, store information, or execute programmed instructions.

– *Intrinsic vs systemic factors*: the computational solution, given the special role assigned to representations, is centered on the description of intrinsic factors that drive the system's behavior. There is no need to specify what kind of measuring devices or effectors are needed, for there may be multiple physical realizations of the same computational scheme, but only the intrinsic (representational) elements are required for the description of the system's behavior. This indicates a fundamental independence of the system's behavior from its non-representational parts, which can be thought as a number of elements extrinsic to the system. The dynamical solution does not show this independence, because there are no extrinsic factors in the system described. Here, the system itself is the solution to the governor's problem. It is not possible to individuate any distinction between intrinsic and extrinsic factors because the actual work is done collectively by all the interconnected parts of the system. Systemic factors, then, are all the elements relevant for the dynamical description of some phenomenon, to the extent that they specify all the features needed to physically realize the system described.

According to van Gelder, the divergence between the two kinds of explanation reflects fundamental characteristics of the different models employed in the respective theories. In the following sections we show that this proposal is unsatisfactory, for the difference between dynamical and computational models is not so well defined as in van Gelder's view. For this reason, we argue that our notion of style of explanation could be more useful to grasp the actual conceptual difference between representational and non-representational theories of cognition.

1.3 Dynamical Systems, Time and State Space

Van Gelder proposes that the contrast between computational and dynamic explanations of cognition, conceptually described through the example of the centifugal governor, is reflected by a profoundly different treatment of time in their respective models. In computational models, indeed, time is discrete, because computations proceed step by step. Dynamical models, on the contrary, describe the evolution of the system variables in real, continuous time, the same time-magnitude we use (as van Gelder says) for modeling any physical system. As mentioned above, this is a rather problematic point of van Gelder's proposal. The problem is that it is not true that the class of continuous-time systems exhausts any possible dynamic system. We can have discrete time (and space) dynamical systems, and this means that a description of computational systems as dynamical systems is not precluded.[1]

A general, informal definition of a dynamical system (Giunti 1995; Pinna 2017) is the following. A dynamical system (DS) is a mathematical structure $DS = \langle M, (g^t)_{t \in T} \rangle$ where:

1. T represents *time*, or the set of durations of the system. T may be either the integers or the reals (the entire sets or just their nonnegative portions);

[1] See Beer (1998) for a reply to van Gelder (1998) on this point.

2. M is a nonempty set that represents the *state space*, i.e. the set of all possible states through which a system can evolve;
3. $(g^t)_{t\in T}$ is a family of functions that represents all the possible state transitions of the system. Each element g^t of this family is a function from M to M that represents a *state transition* (or t-advance) of the system, i.e. g^t tells us the state of the system at any time t, assumed that we know the state of the system at the initial instant t_0.

Let x be any state of the system. The family of functions $(g^t)_{t\in T}$ must satisfy two conditions:

a. for any x, $g^0(x)$ maps x to itself;
b. the composition $g^t \circ g^w$ of any two functions g^t and g^w must be equal to the function g^{t+w}, i.e. if x is an arbitrary initial state, the state of the system reached at time $t + w$ is given by $g^t(g^w(x))$.

Depending on the structure of the time set T and the state space M, it is possible to describe four main types of dynamical system:

(a) Continuous time and state space: both the time set and the state space are the set of the real numbers. Systems specified by differential equations and many kinds of neural networks are examples of dynamical systems of this type.
(b) Discrete time and continuous state space: the time set is the set of natural numbers and the state space is the set of real numbers. Examples of this kind are many systems specified by difference equations.
(c) Continuous time and discrete state space: the time set is the set of real numbers and the state space is the set of natural numbers. This is probably the less interesting case. It is, anyway, simple to construct trivial models of this type of dynamical system.[2]
(d) Discrete time and discrete state space: the time set is the set of natural numbers and the state space is a finite or a countably infinite set. Examples of this latter kind are cellular automata and Turing machines.

The possibility of having dynamical system of type d is a major weakness for van Gelder's position on the contrast between dynamical and computational system. Indeed, it is not possible to characterize this contrast as a matter of a fundamental difference between kinds of models, because we can describe both computational and dynamical (in van Gelder's sense) models in the same theoretical frame (the dynamical systems theory). Hence, it is more appropriate to

[2] An example is a dynamical system DS in which any state transition moves the system to a fixed point. Let the time set T of DS be the set of reals, and its state space M be any finite or countably infinite set. Let $y \in M$ be a fixed point, that is to say, a state y such that, for any t, $g^t(y) = y$. The state transitions of DS are then defined as follows. Let $y \in M$; for any $t \neq 0$, for any state $x \in M$, $g^t(x) = y$; g^0 is the identity function on M.

refer this contrast to some typical assumptions of computational cognitive science that are rejected by the dynamical approach to cognition. These assumptions are linked to the role of representations in cognitive explanations (as shown in Sect. 1.2), but computationalism is not necessarily committed to them (see Sect. 3).

2 Development by Design vs Collective Influence

Now we present two research examples that make manifest the explanatory power of dynamical models with respect to classic approaches. The first example consists in the explanation of the development of stepping movements in toddlers given by Thelen and Smith (1994).

Newborn infants spontaneously produce stepping movements when held upright. These movements disappear at 2 months to reappear at 8–10 months, when the child is able to support his own weight. This behavior is generally explained on the basis of a set of genetically encoded developmental instructions (*development-by-design argument*). Thelen and Smith argue that this explanation is not satisfying, for it does not answer some relevant questions such as: why do infants walk when they do? What are the necessary and sufficient conditions for the appearance of new stepping patterns?

They propose a different explanation of the same phenomenon on the basis of the observation that, in the period when infants do not spontaneously produce coordinated stepping movements, they will produce them *if they are held upright on a motorized treadmill*. In this situation, infants are able to compensate for increasing and decreasing of the treadmill speed, and also to make asymmetrical adjustments of leg movements in a treadmill with two belts moving at different speeds. It seems, then, that the treadmill replaces spring-like leg dynamics that occur naturally in mature locomotion, governed by proprioceptive information available to the central nervous system.

> The point here is that without a context, there's no essence of leg movements during the first year. Leg coordination patterns are entirely situation-dependent [...]. There is, likewise, no essence of locomotion either in the motor cortex or in the spinal cord. Indeed, it would be equally credible to assign the essence of walking to the treadmill than to a neural structure, because it is the action of the treadmill that elicits the most locomotor-like behavior.
>
> [...]
>
> Locomotor development can only be understood by recognizing the multidimensional nature of this behavior, a multidimensionality in which the organic components and the context are equally causal and privileged. That is, while neural and anatomical structures are necessary for the expression of the behavior, the sufficiency of a behavioral outcome is only completed with the task and context (Thelen and Smith 1994, pp. 16–17).

Contrary to the development-by-design argument, this explication takes into account environmental/contextual factors and gives them a previously neglected explanatory role. Hence, it becomes possible to give a different and more satisfactory explanation of a phenomenon previously thought to be governed only by an inner developmental clock.

We can recognize in the development-by-design argument the main characteristics of computational cognitive science, like the fundamental role attributed to inner representations and the algorithmic based style of explanation. In this argument, indeed, behavioral changes are considered as the expression of genetically encoded instruction that represent movement patterns, and the step by step fulfillment of those instruction may be easily associated to an algorithmic execution.

On the other side, in the dynamical explanation of the same phenomenon, the attention is focused on the global evolution of the system, rather than on the encoding of local neural patterns. Shifting the attention toward systemic properties means giving a completely different weight to contextual factors, such as changes in bodily and environmental features, that in inner-centered explanations are at best considered as marginal elements, if not epiphenomena.

The second example is Smith and Thelen's dynamical treatment of the famous A-not-B task (Smith and Thelen 2003). The experimental design of this cognitive task, first described by Piaget (1952), is the following. A child is positioned in front of two boxes, A and B. The experimenter hide a toy, by which the child is attracted, inside box A. The child, then, is allowed to reach the box and take the toy. This first trial is repeated several times, until the experimenter, seen by the child, moves the toy from location A to B. At this point, 8 to 10 month-old children make a search mistake, looking for the toy inside the wrong box. This error disappears when children are about 12 month-old. Piaget explained this phenomenon by hypothesizing an innate development of children's ability to recognize the independence of objects from their own actions. Similar development-by-design arguments have been proposed after Piaget's suggestion (Bremner 1978; Diamond 1998; Munakata 1998), all positing a single cognitive improvement, taking place at about 10–12 month, that allows children to accomplish the task.

Smith et al. (1999) collected data from several version of the A-not-B experiment, where they manipulated various parameters: the delay between the hiding and reaching phases, the number of A trials before the B trial, the presence/absence of visual cues, the direction of children's gaze and their posture with respect to the location of the boxes. They found that none of the previously proposed explanations were able to account for the data. According to them, the main mistake of these proposal is that they all look for a single cause through which we could explain the cognitive error, but there is no single cause. On the contrary, a good explanation should consider the role of several parameters that collectively influence children's performance.

Developing the same line of research, Smith and Thelen (2003) proposed an explanatory model where children's possible choices to search in location A

or B are considered as attractors in a movement planning field, whose shape may be influenced by different kinds of inputs (a task input, a specific input and a memory field). Through the manipulation of these inputs it is possible to predict children behavior during experimental sessions, making the error come and go. This means that this error is not due to the children's lack of some cognitive capacity, suddenly acquired at about 12 months, but is better explained by resorting to contextual factors that collectively influence children's behavior.

In the case of the explanations of the A-not-B error, the analogy between development-by-design arguments and the classic computational approach is more subtle. The point, here, is that behavioral change is the result of some single intrinsic property of the cognitive system, be this property described at some (higher) functional or (lower) neurophysiological level. Similarly, in classic computationalism[3] the functional role of representations is considered as an intrinsic property of mental symbols that mediate for those representations. In the dynamical approach, on the contrary, the relevant systemic properties cannot be identified with some single part of the system, but are the result of the collective work of the magnitudes that govern system behavior.

The characteristics of dynamical explanations seen above, seen in contrast to computational cognitive science, are expressly considered as central points of the dynamical approach to cognition by its proponents (Smith and Thelen 1993; Thelen and Smith 1994; Kelso 1995; Port and van Gelder 1995; Tschacher and Dauwalder 2003). Tschacher and Dauwalder (2003) summarize these points in five tenets:

1. Functional ('intentional', 'goal-directed') cognition is not a single, elementary attribute of the mind [...]. Instead, a *synergy*, i.e. a working together of multiple simple processes is proposed as the basis of cognition and action.
2. [...] To understand the mind one must not focus exclusively on the mind. Instead, cognition is *embodied* (the mind has a brain and a body).
3. [...] To understand the mind one must not focus exclusively on the mind. Instead, cognition is *situated* (i.e. the mind is embedded in environmental constraints and is driven by energetic gradients).
4. A fourth conviction [...] is the *interactivist* conviction. The circular causality which synergetics conceptualizes to be at the heart of self-organizing systems permeates not only the mind, but also the world. [...]
5. The fifth conviction is [...] that self-organized dynamics may help explain *intentionality*. Intentionality and the questions concerning consciousness become topics demanding explanation as soon as conviction (1) is put forward (Tschacher and Dauwalder 2003, p. ix).

The authors here do not individuate the main features of the dynamical approach to cognition in a peculiar treatment of time in the *models* used, as van Gelder does. Rather, their focus is on the conceptual aspects that should be included and explained through the new approach. These aspects define a specific set of target phenomena – very different from the one traditionally considered by

[3] See, e.g., Fodor (1980).

classic cognitive science – on which the dynamical approach should concentrate its analyses. Dynamical models are considered as the most promising conceptual tools in order to deal with all of these characteristics of cognition.

Tschacher and Dauwalder's account of the dynamical approach is more generic than van Gelder's one, for it does not define any specific formal condition that dynamical models of cognition should satisfy. This account, on the one hand, is able to include all the types—(a) to (d)—of dynamical systems listed in Sect. 1.3, while some of them are in principle excluded in van Gelder's proposal. However, on the other hand, this genericity may be a source of confusion with respect to the models that should be employed in dynamical cognitive science. In particular, it seems that, in order to answer to Tschacher and Dauwalder's "convictions", the restriction to dynamical models (of any type) remains unjustified. In the following sections, we will show some proposals where computational models are used in a theoretical framework which is based, as well as the dynamical approach, on an explicit anti-representationalist view.

3 An Anti-representational Interpretation of Computationalism

At the end of Sect. 1.3 we assumed, with no further explanation, that computationalism is not necessarily committed with representationalism, even if it is associated with a vast number of influential representationalist theories. To justify our view we present in this section Wells' proposal of using Turing's theory of computation to formalize Gibson's concept of *affordance*.

Andrew Wells proposes a specific interpretation of Turing's theory of computation which he intends to be more faithful to Turing's original position (Wells 1998, 2005). This view, that he calls *ecological functionalism*, starts from the recognition that a Turing Machine (TM)[4] is the model of a *real cognitive phenomenon type*, namely the one consisting of *a human being which carries out computations with the aid of paper and pencil*. This first move allows the identification of the tape of a Turing machine as an *external environment* (corresponding to the sheet of paper used by a human computer) rather than an

[4] The main components of a TM are the following:

1. a finite automaton (Minsky 1967; Wells 2005) consisting of
   - a simple input-output device that implements a specific set of instructions (machine table);
   - an internal memory that holds only one discrete element at each step (internal state) and
   - an internal device (read/write head) that can scan and change the content of the internal memory.
2. An external memory consisting of a tape divided into squares, potentially extendible in both directions ad infinitum;
3. an external device (read/write/move head) that scans the content of a cell at a time and allows the finite automaton to work on the memory tape.

internal memory. The latter identification is, indeed, a misinterpretation that is typically made by computational functionalists, conceptually connected with the general view of cognition as algorithmic transformation carried out on purely *internal* representation. But, if we turn back to Turing's original position, we no more need to restrict computationalism to this internalistic view of cognition. The symbols written on the tape of a TM may well not only represent but actually *be* external objects bearing cognitive meaning for the subject.[5] According to Wells, this interpretation makes the TM's formalism appropriate for giving a formal account of the notion of *affordance*.

Wells' ecological functionalism, in fact, grounds its roots in Gibson's ecological psychology, whose central point is the concept of affordance (Gibson 1966, 1977, 1979). Although the term 'affordance' refers to a rather technical notion, it is by no means simple to give it a clear-cut definition. Gibson himself gives to the term a deliberately vague meaning, such as in the following passage:

> The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, but the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment (Gibson 1979, p. 127).

In the same page, Gibson specifies the concept by giving a concrete example:

> If a terrestrial surface is nearly horizontal (instead of slanted), nearly flat (instead of convex or concave), and sufficiently extended (relative to the size of the animal) and if its substance is rigid (relative to the weight of the animal), then the surface affords support (Gibson 1979, p. 127).

An affordance, then, is a resource, an aid or an obstacle, offered by the environment to the perceptual space of an animal. An object, a ball for example, may have the affordance of 'graspability' if it is a baseball or 'kickability' if it is a football. It seems clear that the concept of affordance implies a special kind of relation between an animal and the environment. Affordances, in Gibson's theory, are *directly perceived*, i.e. their properties must be specified in stimulus information, without resorting to any kind of internal representation. An animal may also fail in the recognition of such properties, namely it could need a learning phase in order to be able to detect an affordance. The concept of affordance, hence, establishes a special link between perception and action, because in Gibson's theory perceiving something, i.e. detecting some affordance in the environment, corresponds to figuring out an opportunity for action – standing up, grasping, kicking etc.

[5] This view is also consistent with Wilson's proposal of a *wide computationalism* (Wilson 1994).

Gibson claims that his theory poses a challenge to the traditional distinction between subject and object in cognitive explanations. The following quotation attests the richness of the notion of affordance and also justifies why it is not simple to give it a precise definition:

> [A]n affordance is neither an objective property nor a subjective property; or it is both, if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer (Gibson 1979, p. 129)

3.1 Formal Accounts of Gibson's Theory

Given these premises, it is not surprising that scholars have found difficulties in finding a suitable formal model which could reflect the richness of the notion of affordance.

In Wells (2002) the Turing machine formalism is used to construct a model of affordance as an alternative to the models proposed by Shaw and Turvey (1981); Turvey (1992) and Greeno (1994).

Wells analysis starts from the distinction of six main features of the concept of affordance:

- affordances are relational, i.e. they are 'predicated of two or more things taken together' (Wells 2002, p. 144);
- affordances are facts of the environment and facts of behavior;
- a set of affordances constitute the niche of an animal, as distinct from its habitat. The term 'habitat' refers to where an animal lives, while a niche represents a complex relationship among affordances in the environment;
- affordances are meanings, i.e. in Gibson's psychology meanings are perceived directly and are independent of the observer;
- affordances are invariant combinations of variables. This is a central point, for it sets theoretical basis for constant perception and for an explanation of animal evolution, viewed from this stance as an adaptation to constant perceptual variables through which the nature offers opportunities for behavior to a perceiving organism. Wells remarks also that the view of affordances as invariants opens up the possibility to have affordances of different order, because a combination of affordances represents a second order affordance an so on;
- affordances are perceived directly, i.e. they do not need to be mediated by internal representation as, for example, perceptions in the symbolic approach.

Wells then turns to a thorough discussion of three models that have been proposed to formalize the concept of affordance.

Shaw and Turvey (1981) assume that a fundamental notion to understand Gibson's psychology is the concept of *duality*[6] between affordances (as features of the environment) and effectivities (as features of animals). Affordances and effectivities in this account represents *duals*, and then there must be a law which transforms an *affordance schema* into an *effectivity schema*. Informally, the concept of affordance is defined this way: 'an object X affords an activity Y for an organism Z on occasion O if and only if there exists a duality relation between X and Z'. The corresponding affordance schema is: $(X, Z, O|X \diamond Z) = Y$ (where the symbol $\diamond$ indicates a relation of compatibility) which is read as "X, Z and O, given the compatibility between X and Z, equal Y" (Shaw and Turvey 1981, p. 387). By applying the law of transformation to this schema we obtain its dual, namely the effectivity schema: $(Z, X, O|Z \diamond X) = Y$ whose interpretation is 'an organism Z can effect the activity Y with respect to object X on occasion O if and only if there exists a duality relation between Z and X'.

Shaw and Turvey used *coalitions* of four categories of entities (bases, relations, orders, values) in order to explain how the basic relation of duality manifests itself in an ecosystem. Coalitions should permit to study the conditions in which the ecological laws connecting duals of affordances/effectivities hold in nature at different levels (*grains*) of analysis.

Wells moves two main criticisms to Shaw and Turvey's model. The first problem is that the formalization they use does not permit to distinguish between syntactic duals and substantive duals: "Syntactic duals can be created by stipulative definition but substantive duals depend on the prior existence of deeper relationships although they will also have syntactic expressions" (Wells 2002, p. 149). According to Wells, Shaw and Turvey's model allows us to infer the existence of a substantive duality only by using a circular argument, i.e. only through the previous stipulation of a syntactic duality. However, Wells' criticism may not be well addressed, because Shaw and Turvey do not seem to start from a syntactic, but from a substantive (or, better, semantic) duality between affordances and effectivities, in the sense that the relation between these two concepts is intended from the beginning (i.e. from Gibson's characterization itself) as a duality.

A second critical argument Wells moves to Shaw and Turvey's model seems more compelling. He contests that the explanation of ecological laws in term of coalitions of entities creates an infinite regress of levels of analysis, because the model permits a potentially infinite multiplication of levels and it is not clear when and why we should stop assuming the existence of a finer grained level.

[6] The authors define the concept of duality as follows: "[A] duality relation between two structures X and Z is specified by any symmetrical rule, operation, transformation or 'mapping', T, where T applies to map X onto Z and Z onto X: that is, where $T(X) \rightarrow Z$ and $T(Z) \rightarrow X$ such that for any relation r_1 in X, there exist some relation r_2 in Z such that $T : r_1 \rightarrow r_2$ and $T : r_2 \rightarrow r_1$; hence, $XRZ = ZRX$ under transformation T" (Shaw and Turvey 1981, p. 381). They also highlight the importance of this concept in logic, mathematics and geometry, showing as examples dualities in logic between DeMorgan laws, theorems in point and line geometries, open and closed sets in topology, etc. (Shaw and Turvey 1981, pp. 382–384).

Another attempt to formalize the notion of affordance has been made by Turvey (1992). His strategy is based on an analysis of the prospective control of animal activity – i.e. planning action. From this standpoint, an affordance is defined as "an invariant combination of properties of substance and surface [of objects, Ed.] taken with reference to an animal" (Turvey 1992, p. 174). An affordance may or may not be actualized on a given occasion, but it nonetheless represents a *real possibility* of action. Turvey assumes a realist position, where affordances are substantial properties of objects and there can exist neither *thingless properties*, nor *propertyless things*. Besides this characterization, Turvey suggests that affordances are dispositions and that they are complemented by effectivities. To formalize both the notion of affordance and the notion of effectivity, Turvey uses a *juxtaposition* function that joins two dispositional properties, one of the environment and one of an organism. The join of these properties makes a third property manifest. In Turvey's formalism, if X is an entity with dispositional property p and Z another entity with dispositional property q, $W_{pq} = j(X_p, Z_q)$ where j is a function which conjoins the properties p and q (juxtaposition function) in such a way that a third property r is made manifest. For example, let W_{pq} be a person-climbing-stairs system, and r a manifest characteristic property of the system W_{pq}. If X is a person with a certain locomotive property (property p) and Z is a stair with certain dimensions (property q), then Z affords and X effects climbing if and only if:

1. $q, W_{pq} = j(X_p, Z_q)$ possesses r;
2. $q, W_{pq} = j(X_p, Z_q)$ possesses neither p nor q;
3. neither Z nor X possesses r.

Wells rejects this definition of affordance and effectivity as too restrictive. Let us take W_{pq} as a hand-grasping-ball system. In this case, property p would be a value within a dimensional interval depending from some specific hand span, while property q would be the diameter of a ball. A ball will be graspable whenever property q is identifiable with property p. This means that, in any specific case of ball graspability, property p and q will be represented by the same value, hence there is no reason to require condition (2) to hold for this system, for it clearly possesses both properties p and q.

The third model of affordance that Wells discusses was developed by Greeno (1994). Greeno analyze the concept of affordance, on the background of situation theory,[7] as a conditional constraint:

> As a simple example, consider moving from a hallway into a room in a building. An action that accomplishes that is walking into the room, which has the desired effect that the person is in the room because of the action. The relevant constraint is as follows:
>
> $\langle\langle$walk into the room$\rangle\rangle \Rightarrow \langle\langle$be in the room$\rangle\rangle$.

[7] In situation theory a constraint is defined as a "dependency relation between situation types" (Greeno 1994, p. 338).

> Affordance conditions for this constraint include the presence of a doorway that is wide enough to walk through as well as a path along a supporting surface. [...] Ability conditions for the constraint include the ability to walk along the path, including the perceptual ability to see the doorway and the coordination of vision with motor activity needed to move toward and through the doorway (Greeno 1994, p. 339).

In Greeno's view affordances and effectivities are represented by sets of conditions under which dependencies of situation types are made possible.

According to Wells, the main problem faced by this approach is that, in order to make a certain relation between situations happen, some conditions may not hold absolutely, but be context-dependent. A given situation could involve both positive and negative conditions, for example, in the case of the ability to walk into the room, we can add to the affordance conditions the fact that there should be no invisible glass inside the door frame. But then the treatment of affordances as conditional constraints is not consistent with Gibson's theory, for negative conditions cannot be perceived directly or be identified with meanings.

3.2 Affordances and Effectivities as Quintuples of a TM

Wells' criticism unveils a major weakness that is shared by the three approaches described in the previous paragraph, i.e. the fact that they all use the term 'affordance' as something pertaining to the environment, and the term 'effectivity' as something referring to some animal features. But we have seen that the concept of affordance, in Gibson's own words, "refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment" (Gibson 1979, p. 127).

Wells considers that the Turing machine has in its architecture the possibility of being an adequate model for an ecological psychology:

> Turing's analysis was an ecological one for at least the following two reasons. First, its fundamental objects, people who calculate and the numerals they write on paper, are defined at the ecological scale [...]. Second, the analysis formalized the operations of a relational system consisting of an agent who reads and writes symbols using the structured environment of paper ruled into squares. The system as a whole carries out numerical computations (Wells 2002, p. 160).

As explained at the beginning of Sect. 3 the central point of Wells' externalist interpretation of the TM's architecture is that the tape is considered as an external environment. This makes the TM a (schematic) model of an agent-environment system. Wells proposes that the different components of the TM can effectively be used to model affordance and effectivities:

– the input configuration of a TM's quintuple (i.e. a pair q_i, s_j where q_i is an element of the set Q of internal states and s_j is an element of a finite set S of symbols belonging to the tape alphabet) represents an affordance (A) if

we take q_i and s_j to refer to, respectively, the functional state in which an animal happens to be and an external object it finds in the environment; thus $A : (q_i, s_j)$ stands for 'an animal in functional state q_i perceives an object sj'.
– The output configuration of the machine table of a TM (i.e. a triple (s_k, M, q_r) where s_k is another element of the set S, M is a moving operation and q_r is another element of the set Q) represents an effectivity (E) if we take s_k and q_r to refer to, respectively, an animal's behavior (corresponding to an environmental change) and a new functional state to which the animal moves, while M is an element referring to the animal as well as to its environment, "because it represents a movement of the animal relative to the environment" (Wells 2002, p. 161). Thus $E : (s_k, M, q_r)$ stands for 'an animal performs a behavior s_k , moves toward M and changes its mental state to q_r'.

From this standpoint, the machine table of a TM can be seen as a set of affordances coupled to their relative effectivities. In Gibson's terms, the machine table of a TM individuates a niche:

> The complementarity between animal and environment is captured in the way that the set of instructions relating affordances to effectivities specifies the way that the animal behaves. Turing machines have both structure and dynamics and are thus capable of providing models of the animal, the environment and behavior (Wells 2002, p. 161).

This formalization of affordances and effectivities has also the advantage to make these concept independent from the animal/environment – or, philosophically speaking, subject/object – dichotomy, for affordances and effectivities are formalized in such a way that they include terms which refer to both. Another point worth noting is that this formalization models in a natural way affordances as *invariant combinations of variable*. Indeed, affordances, which are specified by input pairs of TM's quintuples, take their terms from a finite set of internal (functional) states and a finite set of symbols (objects), and each type of combination is associated to an output triple (an effectivity), two terms of which are composed of elements taken from the same sets of those composing the input pair. This characterization of affordances permits, on the one hand, *behavioral flexibility*, because the same object, due to a successive learning and/or adaptation phase, may be associated in different times to a different affordance through a change in the animal's internal state (the first element of the input pair). On the other, it guarantees *behavioral stability*, because there is no possibility for an affordance to be associated to different effectivities in different times without structural changes in the animal's niche (see Sect. 3.3); the same affordance will be constantly linked to the same perception and will constitute the basis for successful adaptation. Indeed, behavioral changes and adaptation may be easily associated in this view to possible *extensions* of a machine table.

Table 1. Niche R: the creature is able to find food at the right side of its environment.

Affordance	:	Effectivity		
$Start_1$	Den : Den	R		$Search_1$
$Search_1$	noF : noF	R		$Search_1$
$Search_1$	F : F	H		Eat_1
Eat_1	F : noF	L		$ComeBack_1$
$ComeBack_1$	noF : noF	L		$ComeBack_1$
$ComeBack_1$	Den : Den	H		$Start_1$

Table 2. Niche L: the creature is able to find food at the left side of its environment.

Affordance	:	Effectivity		
$Start_2$	Den : Den	L		$Search_2$
$Search_2$	noF : noF	L		$Search_2$
$Search_2$	F : F	H		Eat_2
Eat_2	F : noF	R	$ComeBack_2$	
$ComeBack_2$	noF : noF	R	$ComeBack_2$	
$ComeBack_2$	Den : Den	H		$Start_2$

3.3 Niche Evolution: A Scenario

Let us imagine a creature with a very simple behavior. It lives in a unidimensional environment, such as the tape of a TM. Starting from its den, it can explore its environment in search of food; once a piece of food is found, it comes back to its den and restarts the same process. Initially, this creature is able to explore only one side of its environment (e.g. all the space on the right of its initial position), with no possibility to reach the resources located on the other side.

To build a TM model of this simple behavior we have, first, to define a tape alphabet and the set of internal states. Let A : {Den, F, noF} be the tape alphabet, where Den is the initial square whose position is never changed, F indicates the presence and noF the absence of food in a square.

Then, we define the set of internal states. We use four internal states:

- $Start_1$: the initial internal state;
- $Search_1$: the tape's head moves to the right, until a symbol F is found;
- Eat_1: the tape's head replaces the symbol F with noF;
- $ComeBack_1$: the tape's head comes back to the square marked with the symbol Den, and all the process restarts.

Similar to a standard TM, the behavioral rules of this creature are defined by a set of quintuples. The rules depend on a set of input pairs (internal state, symbol read) and the possible outputs consist in a set of triples (symbol written, movement, new internal state). The admitted movements are R (move to the adjacent right square), L (move to the adjacent left square), and H (halt, null movement).

The main limitation of this creature's 'niche' (see Table 1) is that one side of the environment is left completely unexplored. We can imagine a totally opposite behavior, where the creature is able to find food only at the left of its initial position. To model this, it is sufficient to change all movement symbols R with L, and vice versa. We also changed the internal states indexes in order to differentiate this niche from the previous one (see Table 2).

Table 3. Niche L+R: the creature is able to find food at both sides of its environment.

Affordance		:		Effectivity	
$Start_1$	Den	:	Den	R	$Search_1$
$Search_1$	noF	:	noF	R	$Search_1$
$Search_1$	F	:	F	H	Eat_1
Eat_1	F	:	noF	L	$ComeBack_1$
$ComeBack_1$	noF	:	noF	L	$ComeBack_1$
$ComeBack_1$	Den	:	Den	H	$Start_2$
$Start_2$	Den	:	Den	R	$Search_2$
$Search_2$	noF	:	noF	R	$Search_2$
$Search_2$	F	:	F	H	Eat_2
Eat_2	F	:	noF	L	$ComeBack_2$
$ComeBack_2$	noF	:	noF	L	$ComeBack_2$
$ComeBack_2$	Den	:	Den	H	$Start_1$

Now, a useful improvement for this creature's behavior would be the ability to reach food located in both sides of its environment. It is possible to model this niche evolution by just extending niche R with niche L with a slight structural change in the internal states: $ComeBack_1$ will now lead to $Start_2$, and $ComeBack_2$ to $Start_1$ (see Table 3). With this niche evolution (modeled as the extension of a TM's machine table), the creature will now search food to the right side of the environment, come back to its den, search to the left, come back to its den, and so on.

4 Conclusion: Privileged Models vs Styles of Explanation

In cognitive science, a lot of effort has been spent in searching for a *privileged* model of cognition, i.e. something that could in principle be used to describe and explain all relevant processes of cognitive phenomena. Any approach to the study of cognition such as classic computationalism, connectionism, dynamicism, enactivism etc., aim to give a *true* definition of what cognition is and how it could be explained with reference to some privileged model. For example, classic computationalism considers mental phenomena as algorithmic transformations of mental symbols that *represent* all external (environmental) and internal (mental) aspects that are relevant for the description of cognitive processes. In the case of classic computationalism, then, the general features of the models employed in the theory – e.g. their being representational – are considered as essential characteristic of cognitive phenomena – mental states *are* representational in nature.

However, we have shown in this paper that same models can be interpreted in different ways, to the extent that some alleged fundamental features of a model in one interpretation result utterly irrelevant in the other. It seems, for instance, that we are not committed to a representational theory of mind if we want to use a computational model in order to describe some cognitive phenomenon. Wells'

proposal, indeed, clearly shows the possibility to use a computational model for the formalization of a non-representational cognitive theory. Other theories where computational models are used to formalize both representational and non-representational aspects of cognitive phenomena are, e.g., Wilson's *wide computationalism* (Wilson 1994) and Giunti and Pinna's *dynamical approach to human computation* (Giunti and Pinna 2016). In all these examples, computational models are used in the context of non (purely) representational cognitive theories. This contradicts the idea that the choice of a model type directly affects the kind of explanation given by a cognitive theory.

To summarize, we have seen that the contrast between representational and non-representational explanations is not necessarily dependent on the type of cognitive model used by a theory:

a. on the one hand, the mathematical tools used in the context of the dynamical approach to cognition, traditionally viewed as a non-representational alternative to representational/computational cognitive theories, are in fact so general that they can be used for the analysis of many different types of models, including computational ones;
b. on the other hand, there is no insuperable obstacle to the use of computational models in a non-representational framework.

Given these considerations, how can we characterize the difference between the two kinds of explanations, i.e. representational vs non-representational ones? To answer this question, we cannot refer to the type of model used by the theory. The explanatory difference may depend not only on the model choice, but on peculiar differences in the target phenomenon as well. To explain this profound conceptual difference we use the notion of *style of explanation*, which indicates the stance taken by a theory with respect to its explanatory targets (see Sect. 1.2).

Representational and non-representational theories, indeed, aim to the explanation of very different properties of the target phenomenon and/or the cognitive system modeled:

(i) representational theories focus on the inner unfolding of cognitive processes, i.e. they aim at describing and explaining the flux of mental states and processes, with specific reference to internal causal factors, such as intrinsic characteristics of mental symbols;
(ii) in contrast, non-representational theories focus on systemic properties, namely on the properties of a cognitive system as a whole, and not on the activity of some specific part. This kind of explanation is better fitted to address developmental and evolutionary questions that remain unexplained in most representational theories, as well as environmental influences on cognition.

From this descriptive analysis of different cognitive theories, however, we should not draw any normative consequence, in the sense that there are no major obstacles to the construction of hybrid theories, which can be used to explain

internal characteristic of cognitive processes as well as the systemic properties of a cognitive system.

Our analysis suggests a further methodological indication. In the case of complex phenomena like cognition, we should give up searching for a privileged model-type; instead, we should focus on the most promising explanatory style for the target phenomena, and choose our model-type according to its ability to best implement that explanatory style for the phenomenon at stake. So we should ultimately spend our greater effort in the definition of target phenomena, rather than in the attempt to adapt phenomena to our models.

Acknowledgements. This work is supported by Fondazione di Sardegna and Regione Autonoma della Sardegna, research project "Science and its Logics: the Representation's Dilemma," Cagliari, CUP F72F16003220002.

References

Beer R (1998) Framing the debate between computational and dynamical approaches to cognitive science. Behav Brain Sci 21:630

Bremner JG (1978) Egocentric versus allocentric spatial coding in nine-month-old infants: factors influencing the choice of code. Dev Psychol 14(4):346

Diamond A (1998) Understanding the a-not-b error: working memory vs. reinforced response, or active trace vs. latent trace. Dev Sci 1(2):185–189

Fodor JA (1980) Methodological solipsism considered as a research strategy in cognitive psychology. Behav Brain Sci 3(01):63–73

van Gelder T (1995) What might cognition be, if not computation? J Philos 92(7):345–381

van Gelder T (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:615–665

Gibson J (1966) The senses considered as perceptual systems. Houghton Mifflin, Boston

Gibson J (1977) The theory of affordances. In: Shaw R, Bransfordk J (eds) Perceiving, acting, and knowing, toward an ecological psychology. Lawrence Erlbaum Associates, Hillsdale

Gibson J (1979) The ecological approach to visual perception. Houghton Mifflin, Boston

Giunti M (1995) Dynamical models of cognition. In: Port R, van Gelder T (eds) Mind as motion. The MIT Press, Cambridge, pp 71–75

Giunti M, Pinna S (2016) For a dynamical approach to human computation. Logic J IGPL 24(4):557–569

Greeno J (1994) Gibson's affordances. Psychol Rev 101:336–342

Kelso JAS (1995) Dynamic patterns: the self-organization of brain and behavior. The MIT Press, Cambridge

Minsky ML (1967) Computation, finite and infinite machines. Prentice-Hall, Englewood Cliffs

Munakata Y (1998) Infant perseveration and implications for object permanence theories: a PDP model of the AB task. Dev Sci 1(2):161–184

Piaget J (1952) The origins of intelligence in children. International Universities Press, New York

Pinna S (2017) Extended cognition and the dynamics of algorithmic skills. Springer, Cham

Port R, van Gelder T (eds) (1995) Mind as motion. The MIT Press, Cambridge
Shaw R, Turvey M (1981) Coalitions as models for ecosystems: a realist perspective on perceptual organization. In: Kubovy M, Pomerantz J (eds) Perceptual organization. Lawrence Erlbaum Associates, Hillsdale
Smith L, Thelen E (eds) (1993) A dynamic systems approach to development. MIT Press, Cambridge
Smith L, Thelen E, Titzer R, McLin D (1999) Knowing in the context of acting: the task dynamics of the A-not-B error. Psychol Rev 106:235–260
Smith LB, Thelen E (2003) Development as a dynamic system. Trends Cogn Sci 7(8):343–348
Thelen E, Smith L (eds) (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge
Tschacher W, Dauwalder J (eds) (2003) The dynamical systems approach to cognition. World Scientific, Singapore
Turvey M (1992) Affordances and prospective control: an outline of the ontology. Ecol Psychol 4:173–187
van Gelder T (1999) Dynamic approaches to cognition. In: Wilson RA, Keil FC (eds) The MIT encyclopedia of the cognitive sciences. MIT Press, Cambridge, pp 244–246
van Gelder T, Port R (1995) It's about time: an overview of the dynamical approach to cognition. In: Port R, van Gelder T (eds) Mind as motion. MIT Press, Cambridge
Wells A (1998) Turing's analysis of computation and theories of cognitive architecture. Cogn Sci 22:269–294
Wells A (2002) Gibson's affordances and Turing's theory of computation. Ecol Psychol 14:140–180
Wells A (2005) Rethinking cognitive computation: Turing and the science of the mind. Palgrave Macmillan, Basingstoke
Wilson RA (1994) Wide computationalism. Mind 103(411):351–372

The Logic of Dangerous Models

Epistemological Explanations for the Incomprehensible Existence of Conspiracy Theories

Selene Arfini(✉)

Department of Humanities - Philosophy Section, University of Pavia, Pavia, Italy
selene.arfini@unipv.it

Abstract. In this paper I aim at examining the use of model-based reasoning for the evaluation of particular explanatory theories: Conspiracy Theories. In the first part of the paper I will take into account the epistemological relevance of Conspiracy Theories: I will discuss their explanatory reach and I will propose that they give their believers the illusion of understanding complex socio-political phenomena. In the second part of the paper I will examine two traditional questions regarding Conspiracy Theories brought forward by the epistemological literature: can Conspiracy Theories ever describe possible conspiracies? Are they in principle non-credible? I will argue that these questions bring forward an epistemic and ontological paradox: if a Malevolent Global Conspiracy (term coined by (Basham 2003)) actually existed, there would be no Conspiracy Theory about it, and if a Conspiracy Theory brings forward details about the existence of a Malevolent Global Conspiracy, there is probably no such conspiracy. I will also specifically address the epistemological issues of discussing the definition of Conspiracy Theories by considering them explanations that brings out the Illusion of Depth of Understanding (term coined by (Ylikosky 2009)) and, with this concept, I will also give reasons to justify their cognitive appeal in the eyes of the lay public.

1 The Epistemology of Conspiracy Theories

The development and use of model-based reasoning in scientific practice has been the core of the debates on models in the philosophical literature of the last decades (Magnani and Casadio 2016; Ippoliti et al. 2016; Magnani and Bertolotti 2017). Questions regarding the use of counterfactual and explanatory reasoning in the creation of material, digital, computational, or mental models have arisen and have been answered with reference to the scientific context, where the production of knowledge has always been the key goal of the research. This approach illustrates and refines our view on the value of models in science, but it does not aim at expanding our knowledge on the epistemological and cognitive role of model-based reasoning in a broad perspective, for instance when used in ordinary situations and lay contexts. Moreover, an investigation on the adoption of model-based reasoning for ordinary explanations is now more than needed since modeling is currently also widely recognized as one of the fundamental cognitive capacities of humans, as well as pattern matching and manipulating the environment (Ifenthaler and Seel 2012; Bransford 1984; Rumelhart 1980). Thus, it seems reasonable

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 41–54, 2019.
https://doi.org/10.1007/978-3-030-32722-4_3

to extend the investigation of the epistemological and cognitive role of model-based reasoning beyond the rigorous use of it in scientific contexts, even considering cases where it fails to provide a widening of the reasoner's knowledge or understanding.

In order to develop this side of the research starting from its most extreme examples, in this paper I aim at examining the use of model-based reasoning in the evaluation of particular explanatory theories: conspiracy theories (henceforth, CTs for short).

Indeed, from an epistemological point of view CTs are extremely fascinating.[1] On the one hand, they represent a peak of mistrust against official channels and authorities (which mainly derives from the extended use of counterfactual explanations) that drive lay reasoning to exceptional levels of skepticism (Clarke 2002). On the other hand, since most CTs are inherently unfalsifiable (Keeley 1999; Basham 2003; Dentith 2014), they also demand a high level of acceptance and blind trust from their believers – even forms of utter dogmatism – about the sophisticated explanatory models of society they create from few core assumptions and hypotheses. Just this clash of contradictory epistemological tendencies should be regarded as bizarre, without even considering the contents of the most accepted CTs. However, if we do consider the contents of some of the more broadly accepted CTs, we would be even more bewildered: the flat Earth hypothesis, the theories that support anti-vaccines movements, and the conjectures on the cover-ups of various celebrities' disappearance as Lady Diana, Elvis, JFK; they all seem too wild and speculative to be taken seriously. Even so, the number of people who believe them is way too high to not raise epistemological wonder.

So, to deal with this multifaceted topic I will proceed by dividing the paper in two main parts. In the first part of the paper I will take into account the epistemological relevance of CTs with reference to issues of social epistemology and psychology, in particular, explanatory reasoning and understanding. I will discuss the explanatory reach of CTs, discussing the philosophical and epistemological investigation of them as lay forms of reasoning and explanations: in particular I will discuss the type of explanatory reasoning CTs develop and why it may foster what Ylikosky (2009); Keil (2003) call *the illusion of depth of understanding* in their believers. I will also analyze the general definition of CTs (and some of its variations), considering the epistemic difference between CTs and other hypothetical theories regarding the existence of potential conspiracies.

In the second part of the paper I will address the models that some authors have put forward to understand CTs and to discriminate between epistemically warranted and unwarranted CTs. In particular, I will take into account the traditional questions regarding CTs brought forward by the epistemological literature (Keeley 1999, 2003; Basham 2003):

A (the ontological problem) can the CTs ever describe possible conspiracies?
B (the epistemological problem) are CTs in principle non-credible?

[1] In this paper the words epistemological and epistemic are meant to address topics and concepts not strictly related to the area of philosophy of science (which the etymology of the two words suggests), but they are used following the current English meaning, reported by the Oxford Dictionary of English (Stevenson 2015, p. 425) as "relating to the theory of knowledge, especially with regard to its methods, validity, and scope, and the distinction between justified belief and opinion.".

I will argue that the ontological problem and epistemological problem of CTs (and the Keeley-Basham model that emerges from the analyses of those problems) bring forward a paradox: it claims that if a Malevolent Global Conspiracy (term coined by Basham (2003)) actually existed, there would be no CT about it, and if a CT brings forward details about the existence of a Malevolent Global Conspiracy, there is probably no such conspiracy. Hence, I will deem that those assumptions present CTs as nonsensical in nature, demeaning the questions regarding their potential credibility and the possibility of the conspiracies to which they refer. Finally, after reconsidering the effects of the illusion of depth of understanding, I will also try to give reasons to justify the cognitive appeal of CTs in the eyes of the lay public.

2 The Epistemological Relevance of Conspiracy Theories (and Their Models)

The basics elements of CTs are not hard to find in popular lay explanations for any event that has a major impact on people's life, such as governmental policies, societal crises, civic and even natural disasters. All that is needed is the belief that a particular event or a gradual phenomenon is caused by the willing efforts of an organized group of people that wants to keep secret its intentions (Keeley 1999, 2003; Basham 2003). The belief does not need to be consistent with current scientific knowledge (Goertzel 2010), nor needs to acknowledge actual goals of political parties (York 2017), nor needs to be consistent with other similar CTs (Wood et al. 2012).

The nature and motive of the hypothetical group of people involved is not even that important to discriminate between CTs: some believers in CTs do declare to be unsure about who is part of the conspiracy they believe in, nor if the group of people is from national or international organizations (Brotherton 2015). Reptilians, governmental officials, international organizations, religious cults, "round Earth" teachers: for a believer in CTs a given conspiracy could have many faces, and reasons, to be orchestrated. The only thing that they are certain about is the existence of at least one big, secret, and evil conspiracy. Often the same event is even explained by believers by appealing to different possible causes or conspiracy groups, or even different CTs. For example, an experiment conducted by (Wood et al. 2012) demonstrate that the more participants believed that princess Diana faked her own death, the more they believed that she was murdered. Even more shockingly, as the psychologist Brotherton (2015) reports, people who strongly believe in a particular CT can even accuse other believers and spokepeople that denounce the existence of the same CT to be part of the conspiracy.

For all these reasons, the technically neutral formula "Conspiracy Theory" entails now, for native English speakers, a negative connotation. It does not just describe a theory about a conspiracy: it usually refers to the belief that a really unlikely or a nonexistent conspiracy orchestrated a major hoax to harm large groups of people – as the belief that a race of reptilians secretly controls the human society and economy to the damage of the poorest and most vulnerable citizens.

The lack of internal consistency in the explanations given to justify the existence of the conspiracy, the belief in more or less far-fetched groups of conspirers, and the hyperbolic intentions that CTs believers attribute to them (to do evil acts), make most

CTs a laughing matter for scholars and intellectuals in general. We also know that CTs are very popular in today's news and it seems easy to approach them from a sociological and a political point of view. Thus, why, if not for curiosity regarding the limits of the human imagination, would epistemologists investigate the nature, structure and becoming of CTs?

I can offer some reasons to conduct an epistemological investigation of them: first, because they put forward issues of practical and social epistemology (since the comprehension of the lay reasoning is one of the focus of those disciplines); second, because believers of CTs exploit various forms of explanatory reasoning, that can shed some lights on how explanations work in a lay perspective; third, because the received view of CTs can enlighten us on the theoretical models that support the believers' constructions. So, let us begin by considering the analysis of CTs from a social-epistemological and -psychological point of view

2.1 Why Conspiracy Theories Matter for Social Epistemology and Psychology

The sketchy, sometime far-fetched, stories that compose the explanations given by people who believe in CTs are actually the blueprint for different forms of reasoning (as well as reasoning biases and heuristics, (Brotherton and French 2014)). These forms of reasoning are already studied by social epistemologists and psychologists when they approach lay beliefs and explanations broadly conceived. Moreover, the defiance of expertise and warranted testimony that people who believe in CTs show is considered one of the causes of the growing popularity of anti-scientific stances in democratic states and anti-intellectualism in populistic parties (Goertzel 2010; Jolley and Douglas 2014; York 2017). Thus, the topic of CTs has also recently come under the attention of both social epistemologists and psychologists,[2] who have examined the questions brought out by people who believe in CTs and found interesting patterns of causal reasoning.

First of all, the hypothesis that a conspiracy exists, that is an organized group of people willingly provoked a particular event – in other terms, the hypothesis that an event has an agent-based cause – is not always an illogical proposition to consider. Conspiracies are more or less common in small and large groups of people: a romantic affair can be described as the smallest and more common conspiratorial group to secretly plot against another party; intelligence agencies need to conspire to promote national security; terrorists conspire to attack the interests of their alleged enemies. Not mentioning the fact that some of the most dramatic events in recent history have been attributed to conspiracies, even if already uncovered (or partially uncovered): the one that resulted in Kennedy's assassination, the Watergate scandal, the Al-Qaeda attack during 9/11, and others.

So the problem is not strictly related to the fact that people believe in the existence of a conspiracy. The problem is that the CT mindset is reported to depend on the use and abuse of biased reasoning and heuristics that make some individuals consider the possibility of a conspiracy the *most likely* hypothesis amongst many others. In particular the psychologists Robert Brotherton and colleagues made different studies on the

[2] To mention just some of the most thorough investigations, Cf. Keeley (2003); Coady (2006); Brotherton et al., (2013).

relation between the tendency to believe in CTs and the propensity of falling into particular reasoning biases and heuristics: Brotherton and French (2014) gave details on the strict relation between the tendency of believe in CTs and the susceptibility to the Conjunction Fallacy; Brotherton and French (2015) reported that individuals biased toward seeing intentionality as the primary cause of events (Biased Attribution of Intentionality) are also more likely to endorse CTs; Brotherton and French (2015) showed that there is a relationship between boredom proneness and conspiracist ideation, mediated by paranoia.

So, when talking of believers in CTs, what is psychologically and epistemologically deviant is not the belief that a conspiracy is in place but the fact that the reasoning that leads people to that conclusion as the only possible one (often without giving consideration to explanations provided by official channels) is bent by biased reasoning and faulted explanations. The epistemological and psychological investigation on the etiology of the CT mindset is then of high importance to distinguish the cognitive appeal of forms of explanatory reasoning that are preferred by people who believe in CTs from more useful (if not more rational) tendencies in lay explanations.

CTs are also relevant when discussing topical issues of social and practical epistemology, in particular the concepts of trust, expertise, testimony, and, more generally, forms of communication between the scientific community and laypeople. In fact, reportedly, CTs have struck a chord with a public mistrust of science and government (Goertzel 2010; Douglas and Sutton 2015). They can tip the balance between what Irzik and Kurtulmus (2018) call a *basic epistemic public trust in science*, which drives, for instance, people to consult their family doctor, and what they call the *enhanced public trust in science*, which assures the public that "scientists have taken measures to avoid the kind of possible errors that the public would like to avoid" (p. 18). Indeed, some studies show that the diffusion of CTs have led to misguided public educational policies (Slack 2007), resistance to energy conservation and alternative energy (Douglas and Sutton 2015; van der Linden 2015), and droppings of vaccination rates (Oliver and Wood 2014; Jolley and Douglas 2014). Hence, the epistemic relevance of CTs should be taken into consideration when referring to the changing role of the expert in modern societies and the standards of public epistemic trust. Especially considering that mild forms of conspiracy theorizing are not hard to find in every category of people and, when closely analyzed, can hardly be considered completely irrational. More than that, a now popular tendency in the philosophical literature is to defend conspiracy theorizing from accuses of utter irrationality and unreasonableness.

2.2 Explanatory Reasoning and Creative Audacity

The unexpected philosophical (and psychological, cf. (Brotherton 2015)) defense of the reasonableness of CTs derives from two sources of trouble.

The first reason is connected to recent psychological studies that have revealed that the belief in CTs is not actually a fringe phenomenon. The popular imagine of the typical conspiracy theorist as an isolated, middle-aged, uneducated person – usually also, for some reason, male – would describe the CTs mindset as an outcome of poor environments, low-level education, and paranoia. Unfortunately, this image finds no match with the number of actual people who express heavy curiosity or faith in CTs. A study

conducted by Uscinski and Parent (2014) in particular, found that all these stereotypes do not apply, since they revealed that women are just as conspiracy-minded as men [pp. 82–83]; they found no links between education, or income, and conspiracism [pp. 86–87]; and no reliable connection between endorsement of CTs and age [p. 86]. People who believe in CTs are not even technically isolated: especially in recent years, a number of CTs communities have risen to the attention of the public and even made news in the newspapers (EXAMPLES). Moreover, recent research (Swami et al. 2013) suggests also that if we pit conspiracy theorists against skeptics, the conspiracy theorists are more intellectually adventurous than the others. So, we cannot simply reduce the conspiracy mindset to a problem of a little fringe of society without the power, the intelligence, or intellectual resources to know better.

The second reason of the re-evaluation of the reasonableness of CTs is connected to the philosophical analysis of them as forms of explanation. Indeed, the most concise and simple definition of a CT is that "a CT is an explanation for an event that appeals to the intentional deception and manipulation of those involved in, affected by, or witnessing these events" Basham (2003, p. 91). This description should lead us to discriminate between warranted and unwarranted theories about the existence of possible conspiracies, but it does not. The attempt of determining a theoretical difference between theories as the one that exposed the Watergate case and theories like the one that suggests that the World is secretly dominated by lizard people has been done by different authors with surprisingly disappointing results. Indeed, finding a solid theoretical difference between these theories is not easy at all.

First, because if we investigate the type of inference that lays at the core of most CTs, we are not able to find an inductive or deductive structure. Instead, the inference is usually based on a creative assumption that justify a series of "errant data" (Keeley 1999), which are data unaccounted by the official theory or that would contradict it if they were true. So, the structure at the core of the CTs can be more accurately described as a creative abductive inference (Magnani 2009, 2017), which aims at finding a reasonable explanation to data that can at first seem not connected to each other, or at the case. Also, this inference aims at creating a new paradigm from which a particular event or gradual phenomenon can make sense. This means that the inferential structure that originates a CT can be described more precisely as what Hendricks and Faye (1999) call a *trans-paradigmatic abduction*. These authors state that:

> New theoretical concepts can be introduced transcending the current body of background knowledge while yet others remain within the given understanding of things. In such cases two paradigms are competing and the abduction is then dependent upon whether the conjecture is made within the paradigm or outside it. [p. 287]

This concept it is usually helpful to describe the reasoning that stands on the core of techno-scientific advancements and creative scientific thinking. It even brings understanding on how serendipitous discoveries have been made in the history of science (Arfini et al. 2018). So, if we use the idea of trans-paradigmatic abduction to understand how CTs can be reasonably distinguished from theories on actual conspiracies, it is not enlightening at all. Indeed, it makes even more clear the fact that the explanations

brought forward by CTs believers are seemingly more creative, even more explanatory than the ones that emerge from official records (as also argued by Keeley (1999); Basham (2003); Keeley (2003)). CTs seem to explain more because they account for more data than the theories taken into account by the official reports, and they bring forward more creative hypotheses. From an epistemological point of view, simplicity and consistency with background knowledge are the only missing qualities for the explanatory reasoning creators and believers in CTs use. Since CTs are mostly lay explanations, they are not even terrible ones.

Nevertheless, an argument can be raised on the kind of understanding (or illusion of understanding) CTs can bring to the agents who believe them.

2.3 The Illusion of Depth of Understanding and Weirdly Catchy Explanations

The Illusion of Depth of Understanding describes people's tendency to overestimate the detail, coherence, and depth of their understanding. The name of this effect has been coined by Keil (2003), who, with some colleagues (Rozenblit and Keil 2002; Mills and Keil 2004) experimentally investigated the influence of this effect and found that most people are prone to feel that they understand the world with greater detail, coherence, and depth than they actually do. The explanation to this tendency is found in the fallibility of the metacognitive ability that inform us of having understood something: the *sense of understanding*. The sense of understanding is a special kind of feeling of knowing which is highly fallible (Grimm 2009) and can be easily confused with what it should indicate: actual understanding. Indeed, understanding and the sense of understanding are only flimsily related: sometimes the sense of understanding can lead the agent to overestimate her understanding of something and sometime it can lead to underestimate it. Usually the former is the case, but not always.

The Illusion of Depth of Understanding is important if discussed in relation to the compelling nature of CTs explanations, because the sense of understanding occurs primarily for explanations if compared with facts, procedures, and narrative (Rozenblit and Keil 2002; Keil 2003). The sense of understanding plays a similar role to the feeling of knowing: it should inform the agent about when she has enough data to say that she understand something. Usually, it leads the agent to overestimate her understanding because in ordinary circumstances having a comprehensive understanding of something is impractical: having first the sense of understanding misleads the agent to think about understanding as an on-and-off phenomenon instead of one that comes in degrees (Ylikosky 2009) and when one reassures herself that she understands something, she can act about it. Basically, quicker the explanatory reasoning kicks-in the sense of understanding, quicker the agent feels that she needs no more data, and she can act about what she thinks she understood.

A way to partially explain the popularity of CTs, I believe, is by considering the hypothesis that they quickly provide reasons to kick-in the sense of understanding in people with different background knowledge and epistemic goals. Actual understanding is a condition that involves different stages (Ylikosky 2009) and some authors argue that it necessitates having an internal mental model of the object of understanding (Waskan 2006). CTs do not provide their believers the conceptual tools to build an internal model for what they defend: if it was the case, the believers in a CT that suggests that Lady D.

faked her own death would not endorse a theory that claims that she was murdered by order of the queen (Wood et al. 2012). The internal models that the two theories provide would collide since they are inconsistent with each other. But the sense of understanding can still be kicked-in by the two CTs: it is a feeling, not a conceptual endorsement. The conceptual endorsement can follow, but it is not the same as proper understanding.

So, a way to distinguish between epistemically warranted and epistemically unwarranted CTs could be found examining the models the they propose and then discuss whether these models can kick-in the sense of understanding of various individuals quicker than the models offered by the official theories.

3 Theoretical Models of Conspiracy Theories

To take into consideration a general but comprehensive model of CTs, I will refer to the analyses conducted by Keeley (1999, 2003) and Basham (2003) in some papers that they separately wrote in the last decades. They mainly ask two questions:

A (the ontological problem) can the CTs ever describe possible conspiracies?
B (the epistemological problem) are CTs in principle non-credible?

To address these problems, they separately and with different aims created an ideal model of CTs, which I will refer below as the Keeley-Basham Model. I should point out that the aims of the two authors were different, even if they do not arrive at completely different conclusions. Keeley (1999) aimed at answering the epistemological problem by presenting an analysis of CTs that permitted to distinguish between warranted and unwarranted CTs (addressing as a side-problem the ontological one). He especially considered the various CTs that emerged after the Oklahoma City bombing. At the end of the paper he concluded that there is no analytical way to distinguish between warranted and unwarranted CTs, and in order to establish the probability that an actual conspiracy is in place, we should give us time to consider the likability of the case addressing the evidence and reaching a consensus with the scientific community.

Basham (2003) started from the analysis conducted by Keeley, focusing instead on the ontological problem. He drew an even more radical picture, starting from even more radical examples: *Malevolent Global Conspiracies* (henceforward MGCs), which can be described as the conspiracies that just the most ambitious CTs describe, in terms of the complexity of the organization, the wickedness of their goals, and the great range of their target manipulations (the CT about lizard-people describes a fair example of a MGC). He starts from evidencing the fact that there is no way to prove that MGCs are impossible and that the usual objections to the likelihood of these conspiracies fail as well. He suggests that the rational dismissing of most CTs, and in particular the ones that describe MGCs, is done merely for pragmatical reasons: if we believe, or rationally leave open the possibility, that MGCs could exist, then a state of utter paranoia (or even worse, dysfunctional skepticism) would take over society. So, he concluded, if it is true that we cannot prove that MGCs are impossible, or even unlikely, and that we cannot distinguish between warranted and unwarranted CTs, it is nonetheless convenient to take seriously only the national and international threats that we can prove right or wrong on a short time period.

Ultimately, I agree with the pragmatical consideration that Basham drew. Nevertheless, I believe there is also a way to recognize in the most radical examples of CTs, so the ones that describe MGCs, a paradox that can actually bring us to consider them unwarranted, ultimately because they describe unlikely conspiracies. So, in the next subsection, I will describe in detail the Keeley-Basham model, what it offers to the picture of CTs and MGCs; in the second subsection I will argue that the picture offered hides an epistemic and ontological paradox and I will then motivate a different quest for epistemologists interested in CTs, by reconsidering the illusion of depth of understanding and its effect on the believers in CTs.

3.1 The Keeley-Basham Model

The first stone of the Keeley-Basham Model was put down by Keeley (1999), who aimed at providing theoretical reasons to consider some versions of CTs unwarranted by definition.

In order to find a way to discriminate between CTs that are unwarranted by definition (henceforth called UCTs) and warranted ones, he first lay down a series of features that are common of all CTs, whether they could be considered warranted or not. He points out three main elements around which a general CT is built up:

> A CT is a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons – the conspirators – acting in secret. Note a few things about this definition. First, a CT deserves the appellation "theory," because it proffers an explanation of the event in question. It proposes reasons why the event occurred. Second, a CT need not propose that the conspirators are all powerful, only that they have played some pivotal role in bringing about the event. They can be seen as merely setting events in motion. Indeed, it is because the conspirators are not omnipotent that they must act in secret, for if they acted in public, others would move to obstruct them. Third, the group of conspirators must be small, although the upper bounds are necessarily vague. Technically speaking, a conspiracy of one is no conspiracy at all, but rather the actions of a lone agent (Keeley 1999, p. 116).

The UCTs instead involve not only these features, but they also:

1. provide an explanation that runs counter to some received, official, or "obvious" account;
2. they support the belief that the true intentions behind the conspiracy are nefarious;
3. and they make use of errant data, which are data unaccounted by the official theory or that would contradict it if they were true (Keeley 1999, p.117).

These added features, as Basham (2003) extensively commented, should have made the conspiracies described by the UCTs extremely unlikely (so, making us addressing not only the epistemological problem, but also the ontological one). Indeed, since UCTs runs counter the official accounts with any means necessary (exploiting errant data), they are usually unfalsifiable. Conspiracy theorists invoke the idea that official explanations are in part or whole deceptions, so they believe they can rationally interpret prima

facie evidence against their accounts as evidence for the conspiracy. Thus, any proof against the CT becomes a proof that a conspiracy is in play. Moreover, the nefarious intent of the conspiracy is problematic if we consider the fact that the conspiracy the believers depict is often too great to be considered actually controllable. Ambitious CTs would require too great a unity and stability of purpose among too many people to be feasible. Furthermore (this argument is suggested only by Basham (2003)), in our society we have trustworthy public institutions of information. If the conspiracy existed, governmental investigations and the free press would eventually encounter evidence of it and effectively sound the alarm. For these reasons, the ambitious CTs necessary belong to the UCT category and they depict unlikely conspiracy, right?

The argument sounds compelling but, Basham argues, is not enough to consider ambitious CTs unwarranted and unlikely. Indeed, the depiction of UCTs that emerges from this analysis has lot in common with the image of a feasible conspiracy which has not been discovered yet. Let's say that there is a conspiracy well-organized enough to secretly make an event of global importance happen. The secrecy would be necessary only if the occurrence of the event was against the moral or deontological norms of a shared community, so we can say that the community itself would consider nefarious the intentions of the conspirators. The public would require a causal explanation for the event, so the well-organized conspirators would need to set up a believable story to cover up their involvement. Then, to make the official story more believable, the conspirators would need to focus the attention of the investigations on particular details and not relevant others, creating errant data.

So, such kinds of conspiracy would be unlikely, but still very much possible. Indeed, a CT regarding this conspiracy would seem of course unfalsifiable; it would describe a level of control and unity of a group of conspirators unthinkable *as far as we know until the conspiracy is discovered*; and it would mean that governmental investigators and public press haven't the right set of data at hand *yet* or, worse, they have interest in keeping the public in the dark. Now, Basham argues that this reasoning can be especially applied if we take into consideration CTs that describe MGCs, which are highly organized conspiracies, with an extended range of manipulations, and extremely evil. So, he ultimately concludes that even if CTs, in particular the ones that describe MGCs, are unfalsifiable, depict an almost incontrollable organization, and do not take into consideration the possibility of having trustworthy public institutions of information, they still describe possible conspiracies, which makes them in turn not unwarranted by definition.

Now, I believe that the Keely-Basham model is a good one to comprehend the basic structure of CTs. The elements that they point out are indeed present in most if not all the CTs, which can be described as theories about conspiracies, which present these features:

1. they propose an explanation for an event
2. they attribute causal agency to a relatively small group of people
3. they propose that this group acts in secret because their intentions are nefarious
4. they propose that this group controls enough people in social, economical, political environments to succeed in their plans but not enough to stop acting in secret

5. they offer an unfalsifiable explanation for the existence of this group, which is supported by many errant data that are not taken into consideration by the official story
6. they assume that the public institutions of information are already controlled by this group or that they neglet its existence

Basham (2003), as well as (Keeley 2003) later, argue that extreme CTs, as the ones that describe MGCs, are not unwarranted by definitions, but they describe still possible conspiracies not yet discovered. Focusing the attention on the fact that these conspiracies, to be considered still possible, would have to be *not yet discovered*, in the next subsection I would present a reason why I believe that the Keeley-Basham model, while it is a good abstract model for CTs and MGCs, if connected to actual instantiations of CTs implies a paradoxical conclusion. Indeed, if a MGC actually existed, there would be no CT about it, and if there is a CT about a MGC, there is probably no such conspiracy.

3.2 The Paradox of Conspiracy Theories

A way to test a model regarding CTs such as the one proposed by Keeley and Basham is to ask how things would work if a MGC would actually exists, and it would have not been yet discovered. To maximize the features that the Keeley-Basham model proposes, we could consider two cases of conspiracies.

On one hand we could consider the possibility that there is an ongoing simplest conspiracy not yet discovered: a love affair. The conspiracy would cause small events that hurt someone, it would involve just two people (the smallest group possible), it would have come up with sketchy cover-stories, and a not-so-nefarious target. The CTs regarding it would look into evidence that are denied by the two conspirators, it would propose an explanation that would account for different errant data that may or may not be connected to the actual plans of the conspiracy. Of course, it would be regarded as an unfalsifiable explanation if believed against the words of the conspirators (the official report).

On the other hand, we could consider a MGC not yet discovered. It would be the most complex conspiracy ever existed and not yet discovered: it would involve lots of people in politics, economics, socially relevant positions; it would have a thought-through cover-story; it would aim at a very nefarious target; it would control any institutions of public information. A MGC usually describes this kind of scenario: the CT that a race of reptilian anthropomorphic aliens secretly controls the high hierarchies of the planet describes a MGC (which is considered to be true by 4% of the people polled in a 2013 report, while a further 7% said they just weren't sure (Jensen 2013)) and checks all of these parameters.

The paradoxical implication of thinking about the possible existence and epistemological warrant for a CT that describes a MGC can be appreciated if we focus on the question: "could there be both a MGC not yet discovered and yet a related CT that describes it in details?" Notwithstanding the fact that the hypothetical MGC would not be impossible, it would be incredible that a CT could describe it, since the conspirators would be so well-organized and well-spread in the high levels of society.

On the contrary, the smallest and silliest conspiracy would be much more likely to be discovered and described in a credible CT. Indeed, in comparison, it would be relatively easy to discover and describe in detail the smallest conspiracy, while a MGC would entail such a complex organization that, if discovered and described in CT, our entire reality would be put into question. In few words, if such gigantic conspiracy existed, we would not have a clue about it. So, in a way, if an uncovered MGC actually existed there would be no CT about it; since some CTs present the very detailed hypothesis that a MGC actually exists, there is probably no conspiracy at all.

In fact, we can consider a MGC possible only if we think about a theoretical model of a conspiracy not yet discovered (so without a CT on it). So the actual instantiations of CTs that describe MGCs (the ones who believe in lizard-people, or that climate-change is a hoax, for example) have actually too many details at hand to claim that such extensive and complex conspiracies exist and nonetheless the conspirators let the theorists talk.

Notwithstanding the fact that the theoretical model for a malevolent global CT describes a possible conspiracy, the actual instantiations of MGC theories (as the reptilians ones) are either by definition nonsensical or the conspiracy they describe is so organized to be almost invincible, since it also controls the way it is present the conspiracy to the larger society: as a paranoid crazy story.

4 Conclusion

To sum up my argument regarding the analysis of CTs, I need to take into consideration the three sides of this epistemological puzzle: the Keeley-Basham models on CTs, the relevance of CTs as creative explanations, and the social impact of CTs on lay people (and others).

Starting with the first item of this list, I argued that the Keeley-Basham model about CTs works only in theory. The actual CTs that describe MGCs cannot logically got it right. More than that, we have a paradoxical situation on our hands: if an uncovered MGC actually existed there would be no CT about it; since some CTs present the very detailed hypothesis that a MGC actually exists, there is probably no such conspiracy.

The problem is, of course, that this paradoxical claim cannot lead to a simple solution in terms of how could we present this argument to the people that now believe in ambitious CTs. Indeed, we should still consider the fact that CTs propose highly creative explanations that enact the form of trans-paradigmatic abductions: they invite the believer in CTs to consider a particular event or phenomenon from a different knowledge paradigm. This kind of explanation, of course, offers a paradigm that is simpler than the one that is offered by contemporary science. It stimulates the sense of understanding regarding sophisticated systems of political play, social and economical dynamics, and even national and international organizations. Nevertheless, in our deeply specialized society, storing knowledge on our own instead of trusting different experts is not an option, both for lay people and specialists.

We can have a general knowledge regarding how the scientific method is applied in different areas or a specialized knowledge in a scientific or humanities sector: either way, we need to trust the specialists in other areas of expertise to get a coherent picture

of how the world works, how can we play a part in it, and what kind of knowledge we can achieve from our actual position. Defying this perspective does not lead to adopt a healthily skeptic point of view on some events or phenomena, but it feeds our sense of understanding that is triggered even when we cannot achieve an actual understanding on specialized knowledge. The fact that a person that believes in a CT is more likely to believe in others (even if they are inconsistent with each other, (Wood et al. 2012)) is revealing about how trust is built on feelings that not necessarily go with actual conditions of understanding. The existence of believers in MGC theories reveals how deep the need of getting the big picture out of our specialized world remains (even if the believe in MGCs actually defy the rules of logic) and is unfortunately connected to the illusion of the depth of understanding. For this reason, I believe that the subsequent epistemological literature on CTs should account for how these theories emerge and what kind of epistemic needs they answer to.

Acknowledgements. I am grateful to Tommaso Bertolotti, Lorenzo Magnani, Matías Ostas Vélez, John Woods, and Paul Thagard's valuable comments on the earlier draft. I also want to express my gratitude towards the two anonymous referees, for their crucial remarks and knowledgeable suggestions.

References

Arfini S, Bertolotti T, Magnani L (2018) The antinomies of serendipity. How to cognitively frame serendipity for scientific discoveries. Topoi https://doi.org/10.1007/s11245-018-9571-3

Basham L (2003) Malevolent global conspiracy. J Soc Philos 34(1):91–103

Bransford JD (1984) Schema activation versus schema acquisition. In: Anderson RC, Osborn J, Tierney R (eds) Learning to read in American schools: basal readers and content texts. Lawrence Erlbaum, Hillsdale

Brotherton R (2015) Suspicious minds. Why we believe in conspiracy theories. Bloomsbury Sigma, New York

Brotherton R, French CC (2014) Belief in conspiracy theories and susceptibility to the conjunction fallacy. Appl Cogn Psychol 28(1):238–248

Brotherton R, French CC (2015) Intention seekers: conspiracist ideation and biased attributions of intentionality. PLoS One 10(5). https://doi.org/10.1371/journal.pone.0124125.

Brotherton R, French CC, Pickering AD (2013) Measuring belief in conspiracy theories: the generic conspiracist beliefs scale. Front Psychol, 4(Article 279). https://doi.org/10.3389/fpsyg.2013.00279.

Clarke S (2002) Conspiracy theories and conspiracy theorizing. Philos Soc Sci 32(2):131–150

Coady D (ed) (2006) Conspiracy theories. The philosophical debate. Ashgate Publishing Limited, USA

Dentith MRX (2014) The philosophy of conspiracy theories. Palgrave Macmillan, United Kingdom

Douglas KM, Sutton RM (2015) Climate change: Why the conspiracy theories are dangerous. Bull Atomic Sci 7(2):98–106

Goertzel TL (2010) Conspiracy theories in science: conspiracy theories that target specific research can have serious consequences for public health and environmental policies. Eur Mol Biol Organ 11(7):493–499

Grimm SR (2009) Reliability and the sense of understanding. In: Regt HD, Leonelli S, Eigner K (eds) Scientific understanding: philosophical perspectives. University of Pittsburg Press, Pittsburg, pp 83–99

Hendricks FV, Faye J (1999) Abducting explanation. In: Magnani L, Nersessian NJ, Thagard P (eds) Model-based reasoning in scientific discovery. Springer, Boston, pp 271–294
Ifenthaler D, Seel NM (2012) Model-based reasoning. Comput Educ 64:131–142
Ippoliti E, Sterpetti F, Nickles T (eds) (2016) Models and inferences in science. Springer, Switzerland
Irzik G, Kurtulmus F (2018) What is epistemic public trust in science? Br J Philos Sci. https://doi.org/10.1093/bjps/axy007
Jensen T (2013) Democrats and Republicans differ on conspiracy theory beliefs. Public Policy Polling. http://www.publicpolicypolling.com/polls/democrats-and-republicans-differ-on-conspiracy-theory-beliefs/
Jolley D, Douglas KM (2014) The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS ONE 9(2):e89177
Keeley BL (1999) Of conspiracy theories. J Philos 96(3):109–126
Keeley BL (2003) Nobody expects the spanish inquisition. More thoughts on conspiracy theories. J Soc Philos 34(1):104–110
Keil FC (2003) Folkscience: coarse interpretations of a complex reality. Trends Cogn Sci 7:368–373
Magnani L (2009) The epistemological and eco-cognitive dimensions of hypothetical reasoning. Abductive cognition. Springer, Heidelberg
Magnani L (2017) The abductive structure of scientific creativity: an essay on the ecology of cognition. Springer, Switzerland
Magnani L, Bertolotti T (eds) (2017) Springer handbook of model-based science. Springer, Switzerland
Magnani L, Casadio C (eds) (2016) Model-based reasoning in science and technology: logical, epistemological, and cognitive issues. Springer, Switzerland
Mills CM, Keil FC (2004) Knowing the limits of one's understanding: the development of an awareness of an illusion of explanatory depth. J Exp Child Psychol 87:1–32
Oliver JE, Wood T (2014) Medical conspiracy theories and health behaviors in the united states. JAMA Intern Med 174(5):171–173
Rozenblit L, Keil FC (2002) The misunderstood limit of folk science: an illusion of explanatory depth. Cogn Sci 26:521–562
Rumelhart DE (1980) Schemata: the building blocks of cognition. In: Spiro RJ, Bruce B, Brewer WF (eds) Theoretical issues in reading and comprehension. Lawrence Erlbaum, Hillsdale, pp 33–58
Slack G (2007) The battle over the meaning of everything: evolution, intelligent design, and a school board in dover. Wiley, San Francisco, PA
Stevenson A (ed) (2015) Oxford dictionary of english. Oxford University Press, Oxford
Swami V, Pietschnig J, Tran US, Nader IW, Stieger S, Voracek M (2013) Lunar lies: the impact of informational framing and individual differences in shaping conspiracist beliefs about the moon landings. Appl Cogn Psychol 27(1):71–80
Uscinski JE, Parent JM (2014) American conspiracy theories. Oxford University Press, Oxford
van der Linden S (2015) The conspiracy-effect: exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. Pers Individ. Differ 87(1):171–173
Waskan JA (2006) Models and cognition. The MIT Press, Cambridge
Wood MK, Douglas KM, Sutton RM (2012) Dead and alive: beliefs in contradictory conspiracy theories. Soc Psychol Pers Sci 3(6):767–773
Ylikosky P (2009) The illusion of depth of understanding in science. In: Regt HD, Leonelli S, Eigner K (eds) Scientific understanding: philosophical perspectives. University of Pittsburg Press, Pittsburg, pp 100–119
York A (2017) American flat earth theory: anti-intellectualism, fundamentalism and conspiracy theory. History Undergraduate Publ Presentations 3:2–37

A Pragmatic Model of Justification Based on "Material Inference" for Social Epistemology

Raffaela Giovagnoli(✉)

Faculty of Philosophy, Pontifical Lateran University, Vatican City, Italy
giovagnoli@pul.it

Abstract. Social epistemology presents different theories about the status of shared knowledge, but only some of them retain a fruitful relation with classical epistemology. The aim of my contribution is to present a pragmatic model which is, on the one side, related to the classical concepts of "truth" and "justification", while, on the other side, addressing to a fundamentally "social" structure for the justification of knowledge. The shift from formal semantics to pragmatics is based on a notion of "material inference" embedding commitments implicit in the use of language, that favors the recognition of the social source of shared knowledge.

Keywords: Social epistemology · Truth · Justification · Material inference · Deontic statuses · Deontic attitudes

1 Introduction

We'll present a "social" model for knowledge representation made explicit by a form of "expressive logic" as presented by the American philosopher Robert Brandom in his important work *Making It Explicit*, which can be considered as a plausible alternative to relativism in social epistemology (Brandom 1994). This move means to propose a pragmatic order of explanation that focuses on the role of expression rather than representation. In this context, "expression" means to make explicit in assertion what is implicit in asserting something. A fundamental claim of this form of expressivism is to understand the process of explicitation as the process of the application of concept. According to the relational account, what is expressed must be understood in terms of the possibility of expressing it. Making something explicit is to transform it in premise and conclusion of inferences. What is implicit becomes explicit as reason for asserting and acting. Saying or thinking something is undertaking a peculiar kind of *inferentially* articulated commitment. It shows a deontic structure that entails the *authorization* of the inference as a premise and the *responsibility* to entitle oneself to that commitment by using it (under adequate circumstances) as conclusion of an inference from other commitments one is or can become entitled. To apply a concept is to undertake a commitment that entitles to and precludes other commitments.

Actually, there is a relevant difference between the Wittgensteinian theory of linguistic games and the scorekeeping model. Inferential practices of producing and

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 55–68, 2019.
https://doi.org/10.1007/978-3-030-32722-4_4

consuming reasons are the point of reference of linguistic practices. Claiming is being able to justify one's claims and other claims (starting from one's claims) and cannot be considered as a game among other linguistic games. Following Wilfrid Sellars, Robert Brandom uses the metaphor of the "space of reasons", but he understands it as a "social" concept, i.e. as the space of the intersubjective justification of our assertions (Brandom 1995).

The theoretical points we'll discuss in the following sections to introduce a social account for justification are: (1) a brief presentation of some accounts in social epistemology (2) the basic concept of inference an agent must necessary perform; (3) the kind of normativity implied by discursive agency; (4) the structure of the conceptual content and (5) the dimensions of justification in the game of giving and asking for reason.

Reasons contained in assertions possess a content that is inferentially structured. We conclude that the formal structure of communication gives us the possibility to make explicit this content. From the point of view of a "social" concept of the space of reasons, beliefs, mental states, attitudes and actions possess a content because of the role they play in social "normative" practices (inferentially articulated).

2 Communitarian Epistemology

Social epistemology presents different perspectives concerning the assessment of "social evidence"[1]. We can (I) assess the epistemic quality of individual doxastic attitudes when social evidence is used; (II) assess the epistemic quality of group attitudes or (III) assess the epistemic consequences of adopting certain institutional devices or systemic relations as opposed to alternatives (Goldman 2015; Giovagnoli 2017a, b). The so-called "communitarian epistemology" (Hardwig, Welbourne, McIntyre, Brandom) falls into the first stream and, particularly, maintains that knowledge is "essentially" social.

In order to sketch "communitarian" epistemology, it is important to analyze the notion of "evidence" which is a central notion in the philosophy of science and the sociology of scientific knowledge. John Hardwig criticized the "individual" conception of evidence, according to which we can have good reasons to believe "that p" only if we have evidence to support it and evidence is "anything that counts toward establishing the truth of p (i.e., sound arguments as well as factual argumentation)" (Hardwig 1995, p. 337). But, suppose that the doctor I trust told me that I have been suffering for many years of a rare illness of my foot. He has good reasons for the diagnosis because, due to his experience, he can formulate a reliable judgement by

[1] Evidence is a fundamental notion in the ambit of epistemology and classically refers to the the individual mental states and processes that enable the subject to grasp knowledge in a reliable sense. Beyond the classical philosophy of mind, we can consider knowledge as related to the use of ordinary language, so that it depends from competent speakers who undertake suitable roles in discursive situations. So, evidence depends on the correct use of language to be tested in interactive contexts. But, "social evidence" does not only entails the use of ordinary language from the part of competent speakers, it strongly depends on testimony, namely on the use we make of what is transmitted on social contexts. Testimony is at the center of a lively debate in social epistemology which gives rise to different perspectives (Goldman 2015).

studying the radiographies of my foot and my way of walking. I can be skeptical about this conclusion, because, for example, I feel no pain. Still I have good reasons to trust my doctor judgement. But, do my reasons represent evidence for the truth of the diagnosis? According to individualism, the response is negative, because my reasons to believe the diagnosis does not correspond to the ones of my doctor. The good reasons from my doctor are not sufficient to establish a relationship of trust. They do not become stronger after the expression of the diagnosis. But, according to Hardwig, the "narrow" conception of evidence conflicts with common sense, so we must introduce a "wider" notion of evidence. In standard cases, we trust what our doctor says and therefore our reasons correspond to his reasons. We refer to knowledge of experts and in ordinary life would be irrational to do otherwise, as we are not able to check the truth and accuracy of what we come to know. Sometimes, we examine the credentials of the experts when they are in conflict with their colleagues. However, we are not obliged to always use our own head.

Hardwig proposes to extend the authority of testimony (nota) to knowledge in general (Hardwig 1991, p. 698):

> "belief based on testimony is often epistemically superior to belief based on entirely direct, non-testimonial evidence. For (one person) b's reasons for believing p will often be epistemically better than any (other person) a would/could come up with on her own. If the best reasons for believing p are sometimes primarily testimonial reasons, if knowing requires having the best reasons for believing, and if p can be known, then knowledge will also sometimes rest on testimony."

This thesis is supported by arguments from scientific practice. Scientists form routine "teams" on the basis of testimony and trust. Hardwig makes the example of a physicists' team working on high energy in the early 80 years (Hardwig 1995, p. 347):

> "After it was funded, about 50 man/years were spent making the needed equipment and the necessary improvements in the Stanford Linear Accelerator. The approximately 50 physicists worked perhaps 50 man-years collecting the data for the experiment. When the data were in, the experimenters divided into five geographical groups to analyze the data, a process which involved looking at 2,5 million pictures, making measurements on 300.000 interesting events, and running the results through computers…The "West Coast Group" that analyzed about a third of the data included 40 physicists and technicians who spent about 60 man-years on their analysis."

The research has been published in an article from 99 co-authors, some of them will never know how it reached such number. To produce the data for such an article presupposes that scientists exchange information and that they consider the results from the others as evidence for the ongoing measurements. None of the Physicists could replace his knowledge by testimony with knowledge based on perception: it would require to much life time. This type of "epistemic dependence" is visible also in mathematics, for instance in the proof of De Branges of the Bieberbach's conjecture; a proof that involved mathematicians with very different forms of specialization.

Starting from Hardwig's work, Martin Kusch isolates three epistemological alternatives:

(1) "Strong individualism" according to which knowledge presupposes individual sources of evidence.

(2) "Weak individualism" according to which it is not necessary to possess evidence for the truth of what one believes and to completely understand what one knows.
(3) "Communitarianism" according to which community is the primary source of knowledge. It retains the idea that the agent would have "direct" possession of the evidence, but it breaks with the assumption that such an agent would or could be an individual.

Hardwig could be considered as a communitarian not only regarding epistemology but also philosophy in general. Testimony occupies a space where epistemology meets ethics. If a certain result by an expert provides good reasons to believe that p, it will depend on the perception of the receiver about the reliability of the testimony of the expert, that for its part will depend on an evaluation of his character. Hardwig's work on teams and trust in scientific practice has influenced relevant authors in the field of social epistemology (Galison, Knorr Cetina, Shaffer, Shapin e Mackenzie). Kusch underscores two limitations of his approach (Kusch. First, Hardwig privileges scientific communities, so he does not consider cases of cooperation in ordinary life where testimony plays a crucial role: we trust a lot of public messages without investigating the sincerity and the competence of the source. Second, it is not very clear the way in which Hardwig refers to the evidence of a true belief. Beyond individualism, either it must be explained the nature of the evidence possessed by the teams or the process through which we can proceed a true belief.

Michael Welbourne published the book *The Community of Knowledge* (Welbourne 1993), that represents a valid example of communitarian epistemology based on testimony (nota). He makes a fundamental theoretical move; he considers testimony not as "mere transmission of information" (the so called "say so" which characterizes classical epistemology). Knowledge develops in a community, where it is transmitted according to a certain view of "shared knowledge." To share knowledge means to share commitments and entitlements with others, almost in several standard cases. His theory of "authority" is opposed to the theory of evidence. We do not possess direct evidence for knowledge, because entitlements imply anything can serve as ground for our inferences. Knowledge must be objective and so "social" because we consider it as an external and objective standard for what also others should recognize. Commitments we undertake entail an investigation on the entitlements of the others; therefore we create a dialogical dynamics that generate new shared knowledge. Kush makes an example to clarify this dynamics (Kusch 2002, pp. 59–60):

> "Assume that I claim to know how long it takes to travel from Cambridge to Edinburgh; I tell you, and you believe me and tell me so. In doing so, we agree that we should not consent to anyone who suggests a different travel period, that we shall inform each other in case it turns out that we did not possess knowledge after all, that we shall let this information figure in an unchallenged way in travel plans, and so on. We can perhaps go beyond Welbourne by saying that the sharing of knowledge creates a new subject of knowledge: the community. And, once the community is constituted, it is epistemically priori to the individual member. This is so since the individual community member's entitlement and commitment to claiming this knowledge derives from the membership in this community. The individual knows as "one of us," in a way similar to know how I get married as "one of a couple," or II play football as "one of the team."

But, according to Kusch, Welbourne does not consider the normative ground of testimony, namely the background knowledge. This background represents what agents

concretely share and encloses the important results of previous communities of knowledge from which we inherit them. It should be possible to go beyond the dialogical exchange of reasons starting from commitments and entitlements and to rest on knowledge constituted by testimony through a sort of "institutionalization."[2] In this case, we need a theory of social institutions and social states based on the use of the so-called "performatives" (Austin). The major referents for social epistemology are John Searle, Barry Barnes and David Bloor. But, we recall also the work of Kent Bach, Esa Ikonen, Eerik Lagerspetz and Raimo Tuomela.

Performative testimony moves from the act we do by saying anything and from how it is received by our interlocutor. It is not a matter of the simple "say so" or mere transmission but a process of social construction. A performative testimony does not allow us to consider a state of affairs p, reference and knowledge as discrete, sequential and independent events. For example (Kusch 2002, 65–66):

> "The registrar a tells the couple b that they have now entered in a legally binding relationship of marriage; and by telling them so, and their understanding of what he tells them, the registrar makes it so that they are in a legally binding relationship of marriage. For the registrar's action to succeed, the couple has to know that they are being married through his say-so, and he has to know that his action of telling does have this effect. Moreover, a and b form a community of knowledge in so far as their jointly knowing that p is essential for p to obtain. That is to say, a and b enter into a nexus of entitlements and commitments, and it is this nexus that makes it so that each one of them is entitled to claim that p. The registrar has to use certain formulas (By the power invested in me by the state of California …etc.) bride and groom have to confine themselves to certain expressions (a simple "yes" or "no" will be fine), and each one commits himself or herself, and entitles the other, to refer to p as a fact subsequently. More principally, we can say that "getting married" is an action that is primarily performed by a "we."

The new social status and the knowledge the copule shares is generated by performative testimony, namely by the speech act performed by the adequate authority. The reasons why performative testimony generates knowledge entail two important characteristics of performatives: self-referentiality and self-validity. The act refers to itself because it announces what it will happen and, if performed under the right circumstances, it generates the validity of the reality it creates. The act that creates the new social situation is like a common act performed by the agreement among persons. This act is fragmented and distributed on other speech acts, it is implicit in ordinary practices, like when we greet, we talk about greeting the colleagues we meet or we criticize someone who did not respond to our greeting. All these acts are mostly performatives or include a shared performative. This conclusion is relevant for social epistemology as it often realizes through performatives that are shared and widely distributed. Kusch observes that knowledge is a social state constituted by a shared performative (a declaration that there exists a unique way to possess truth and we call it "knowledge"). Knowledge is a social referent create by the references to it; and these references occur in the testimony like in other forms of dialogue. Dialogue includes to

[2] In this case we derive shared knowledge from the processes and states that enable individuals to create, accept and recognize the norms and institutions they create in suitable social contexts. There are several interesting perspectives related to the debate on Collective Intentionality as a capacity that produces shared knowledge and social evidence.

assert that something is knowledge, to challenge knowledge, to test knowledge. to doubt and so on within a wide range of possible references. Testimony can obtain the status of knowledge because we make direct and indirect reference to it through numerous examples of constative and performative testimony. This direct and indirect reference creates knowledge as a social state.

3 The Basic Concept of "Inference"

Before to introduce a social concept of the space of reasons by reference to Brandom's work, we want to make clear the sense in which we are talking of "normativity" as grounded on linguistic rules. Recalling the Wittgensteinian account about what it means to follow a rule and the fact that we cannot follow a rule "privatim", we focus on the discursive structure which enables us to "keep score" in conversation namely to grasp and master concepts using ordinary language[3]. The content of beliefs and actions is "phenomenalistic" because it is expressed by inferential rules in the sense of material incompatibility. Moreover, the grasp of the conceptual content is possible only using intersubjective pragmatic rules that in some sense "harmonize" the collateral beliefs of the participants. An agent must be able to recognize the correct use of concepts by performing correct inferences, which owe their correctness to the fact that they express precise circumstances and consequences of the application of concepts. This theoretical option means that inference must be considered as "material". For example, the inference from "Milan is to the north of Rome" to "Rome is to the south of Milan" is governed by material properties: the concepts of "north" and "south" make inference correct. It is therefore necessary to grasp and to use these concepts in order to perform correct inferences without the necessity to refer to norms of formal logics.

The conceptual content is inferential in scorekeeping terms:

«What's incompatible with such a conditional, e.g. with **if** p **then** q, is what's simultaneously compatible with its antecedent (p) and incompatible with its consequent (q). In any context where one is committed both to a conditional and to its antecedent, then, one is not entitled to any commitment incompatible with its consequent, and that is just to say that one is also committed to that consequent. Assertional commitment to the conditional **if** p **then** q, in other words, establishes a deontic context within which a commitment to p carries with it a commitment to q – but this is just a context within which p (commitment) implies q. It is in this sense that the conditional expresses the propriety of the corresponding inference, without so to speak, also reporting it, as would the corresponding normative metalinguistic claim» (Rosemberg 1997, pp. 179–180).

The role of logic is to make explicit the material properties of the content of our beliefs. This is a crucial move also for autonomy (Giovagnoli 2004, 2007; Giovagnoli

[3] The functioning of scorekeeping in a language game has been presented by David Lewis (Lewis 1983). Brandon inherits the model but he changes it according to an original account of the inferential structure of conceptual content and its relevance for social interaction. The result of Lewis' model is useful to understand the context dependence of ordinary conversation and this option helps us to grasp in plausible way the nature of the content in the *game of giving and asking for reasons* (to use the Sellarsian expression).

and Seddone 2009), because the agent playing the role of scorekeeper undertakes, at the same time, a "critical" perspective.

A good example of the expressive function of logic is Michael Dummett's question of "harmony" (Dummett 1973, 1991). In short, harmony means that I-rules and E-rules must somewhat harmonize: the Fundamental Assumption that complex statements and their grounds, as specified by their I-rules, must have the same set of consequences. Namely, I- and E- rules must be in harmony with each other in the sense that one may infer from complex statement nothing more and nothing less than that which follows from its I-rules (Murzi and Steinberger 2017).

Dummett maintains that the application of a concept directly derives from the application of other concepts: those concepts that specify the necessary and sufficient conditions that determine the truth conditions of claims implying the original concept. This assumption would require an ideally transparent conceptual scheme embedding all the necessary and sufficient conditions for the application of a concept that, consequently, makes invisible the "material" content of concepts. Let's consider the term "Boche" that applies to all the German people and implies that every German is a rough and violent type especially if compared to other Europeans. In this case, the conditional, that makes explicit the material inferences of the use of the concept (if he is Boche so he is rough and violent), enables an adequate criticism that aims to the acceptance or the refusal of certain material commitments. The introduction of the term "Boche" in a vocabulary that did not contain it does not imply, as Dummett suggests, a non-conservative extension of the rest of language.

The substantive content of the concept implies rather a material inference that is not already implicit in the contents of other concepts used for denoting the inferential pattern from an individual of German nationality to a rough and violent individual. In Brandom's terms:

«The proper question to ask in evaluating the introduction and evolution of a concept is not whether the inference embodied is one that is already endorsed, so that no new content is really involved, but rather whether that inference is one that ought to be endorsed. The problem with 'Boche' or 'nigger' is not that once we explicitly confront the material inferential commitment that gives the term its content it turns to be novel, but that it can then be seen to be indefensible and inappropriate – a commitment we cannot become entitled to. We want to be aware of the inferential commitments our concepts involve, to be able to make them explicit, and to be able to justify them» (Brandom 2000, pp. 71–72).

This thought embeds a profound sense for inferentialism: semantics is "answerable to pragmatics". The start point of this kind of inferentialism is the "doings" of linguistically endowed creatures, in particular their practices of asserting and inferring which, according to Brandom "come as package" (Weiss and Wanderer 2010). The speech act of assertion allows us to advance claims expressed by declarative sentences. Assertion is the primary unit of significance by virtue of its very structure grounded on inferential relations. Brandon follows the proposal introduced by Gentzen to divide the model of the meanings of logical expressions in terms of I-rules and E-rules so that assertion acquirers its meaning by virtue of "a set of sufficient conditions" and a "set of necessary consequences" (nota). He also introduces the set of claims incompatible with it. The primacy of assertion is at the center of Habermas' and Kusch's criticism who prefer to rely (even starting from different

models) to the attitudes of a community as a form of collective intentionality for describing the objectivity of knowledge we can share (Giovagnoli 2001).

4 The Role of Conditionals for Human Discursive Practices

We are not only creatures who possess abilities such as to respond to environmental stimuli we share with thermostats and parrots but also "conceptual creatures" i.e. we are logical creatures in a peculiar way. It is a fascinating enterprise to investigate how machines simulate human behavior and the project of Artificial Intelligence, a project that began meads of the XX century, could tell us interesting things about the relationship between syntactical abilities and language. Brandom seriously considers the functioning of automata because he moves from some basic abilities and he gradually introduces more sophisticated practices, which show how an autonomous vocabulary raises (Brandom 2008). This analysis is a "pragmatist challenge' for different perspectives in analytic philosophy such as formal semantics (Frege, Russell, Carnap and Tarski), pragmatics both in the sense of the semantics of token-reflexive expressions (Kaplan and Stalnaker) and of Grice, who grounds conversation on classical semantics. Conditionals are the paradigm of logical vocabulary to remain in the spirit of Frege's Begriffschrift. But, the meaning-use analysis of conditionals specifies the genus of which logical vocabulary is a species. In this sense, formal semantics is no more the privileged field for providing a universal vocabulary or meta-language.

Starting from basic practices, we can make explicit the rules that govern them and the vocabulary that expresses these rules. There are practices that are common to humans, non-human animals and intelligent machines that can be also artificially implemented like the standard capacities to respond to environmental stimuli. But, it seems very difficult to artificially elaborate the human discursive practices which depend on the learning of ordinary language. In particular, humans are able to make inferences and so to use conditionals because they move in a net of commitments and entitlements embedded in the use of concepts expressed in linguistic expressions. Logical vocabulary helps to make explicit the inferential commitments entailed by the use of linguistic expressions, but the meanings of them depend on the circumstances and consequences of their use. The last meta-language is ordinary language in which we give and ask for reasons and therefore acquire a sort of universality. It seems that, we do not need to apply the classical *salva veritatae* substitutional criterion, as conditionals directly make explicit the circumstances and consequences namely inferential commitments and entitlements possessed by singular terms and predicates (Giovagnoli 2012).

The source of the normativity entailed by conceptual activity is a kind of "autonomous discursive practice" that corresponds to the capacity to associate with materially good inferences *ranges of counterfactual robustness* (Brandom 2008; Giovagnoli 2013; Giovagnoli 2017). In this sense, "modal" vocabulary represented by modally qualified conditionals such as if p then q has an expressive role. Modal vocabulary is a *conditional* vocabulary that serves to *codify* endorsements of material inferences: it makes them explicit in the form of material inferences that can themselves serve as the premises and conclusions of inferences. According to the argument Brandom calls "the modal Kant-Sellars thesis", we are able to secure counterfactual

robustness (in the case of the introduction of a new belief), because we "practically" distinguish among all the inferences that rationalize our current beliefs, which of them are update candidates. The possibility of this practical capacity derives from the notion of "material incompatibility", according to which if we treat the claim that q follows from p as equivalent to the claim that everything materially incompatible with q is materially incompatible with p. So, for example if we say "Cabiria is a dog" entails "Cabiria is a mammal" we are stating that everything incompatible with her being a mammal is incompatible with her being a dog.

Brandom makes a move that is mostly inspired by the dialectic of Hegel and, consequently, collapses semantic content into cognitive content: the contents of concept words (which belong to the realm of Fregean sense) are given by their inferential relations. He does not point out what reality is, but he simply maintains that we know it by using our discursive practices (inferentially articulated).

We acquire concepts during the process of language learning and, because language offers a sort of classification of reality, we also acquire a widely "shared knowledge". Classical philosophy is mostly devoted to investigate the way in which knowledge can be acquired and classified[4]. It seems that the research should be extended beyond the mere exercise of reliable responsive dispositions to respond to environmental stimuli, even though we find very fruitful investigations in natural sciences. The conceptual activity is better clarified by intending the application of a concept to something as *describing* it. One thing is to apply a label to objects, another is to describe them. In Sellars' words (Sellars 1956, pp. 306–307):

"It is only because the expressions in terms of which we describe objects, even such basic expressions as words for perceptible characteristics of molar objects, locate these objects in a space of implications, that they describe at all, rather than merely label".

Human use of concepts corresponds to the capacity to endorse conditionals, i.e. to explore the descriptive content of proposition, their inferential circumstances and consequence of application, which characterize a sort of "semantic self-consciousness": the higher capacity to form conditionals makes possible a new sort of hypothetical thought that seems to appear as the most relevant feature of human rationality. Human rational capacities are so characterized by recognizing premises and conclusions of valid inferences that can represent good reasons for what is asserted.

But, beyond this inferential semantics we must underscore the role of pragmatics namely what we do when we endorse a good inference. We do not simply deny a sentence we do not endorse, but we are able to recognize incompatibility relations originated by the inferential structure of semantic contents Consequently, we can derive a distinction (proposed by Michael Dummett) between "ingredient" content and "freestanding" content. Starting from the fact that we master the use of concept during the process of acquisition of ordinary language we can observe that the former belongs to a previous stage where it becomes explicit only through the force of sentence (query, denial, command etc. that are invested in the *same* content). For example, a child come

[4] Brandom moves from classical thought which generally intends the paradigmatic cognitive act as classifying, i.e. taking something particular as being of some general kind. This conception that originates in Aristotle's *prior Analytics*, was common to everyone thinking about concepts and consciousness in the modern period up to Kant.

to acquire concepts in the interaction with the parents and, in this case, questions are a very good device. The latter can be grasped in terms of the contribution it makes to the content of the compound judgments in which it occurs, consequently, only indirectly to the force of endorsing that content. Therefore, the process of human logical self-consciousness could be thought as developing in three steps:

1. We are able to "rationally" classify through inferences i.e. classifications provide reasons for others.
2. We form synthetic logical concepts formed by compounding operators, standardly by conditionals and negation.
3. We form analytic concepts where sentential compounds are *decomposed* by checking invariants under substitution.

The third step gives rise to the "meta-concept" of ingredient content i.e. we realize that two sentences that possess the same pragmatic potential as free-standing, favoring rational classifications, nevertheless make different contributions to the content (and consequently the force) of compound sentences where they occur as unendorsed components (Brandom 2012). When we substitute one for another, we see that the free-standing significance of asserting the compound sentence containing them can change. We learn how to form complex concepts by applying the same methodology to sub-sentential expressions (singular terms), that repeatedly occur in those same logically compound sentences. This process gives rise to various equivalence classes that can be regarded as substitutional variants of one another. It represents a distinctive kind of analysis of those compound sentences, a sort of hierarchy, because it entails the application of new concepts. Actually, they were not components out of which they were originally constructed. The most impressive result of this kind of research is in the ambit of what Brandom's logics can express: concepts so originated are substantially and in principle more expressively powerful than those available at earlier stages in the hierarchy of conceptual complexity (they are, for instance, indispensable for even the simplest mathematics).

5 The Dimensions of Justification

We master the use of concepts in the process of language acquisition. So, we are here interested in kind of representation of the content of linguistic expressions that goes beyond the mere "mental" *representation of objects*. This is the main reason to discuss Brandom's perspective also in the ambit of social epistemology to present an interesting model for the objectivity of shared knowledge.

The scorekeeping model replaces the Kantian notion transcendental *apperception* with a kind of synthesis based on incompatibility relations. In drawing inferences and "repelling" incompatibilities, a person is taking oneself to stand in representational relations to objects that she is talking *about.* A commitment to A's being a horse does *not* entail a commitment to B's being a mammal. But it *does* entail a commitment to A's being a mammal. Drawing the inference from a horse-judgment to a mammal-judgment *is* taking it that the two judgments represent one and the same object. Thus, the judgment that A is a horse is not incompatible with the judgment that B is a cat. It *is*

incompatible with the judgment that A is a cat. Taking a horse-judgment to be incompatible with a cat-judgment *is* taking them to refer or represent that object, to which incompatible properties are being attributed by the two claims.

The normative *rational* unity of apperception is a synthesis to expand commitments inferentially, noting and reparing incompatibilities. In this sense, one's commitments become reasons for and against other commitments; it emerges the rational critical responsibility implicit in taking incompatible commitments to oblige one to *do* something, to update one's commitment so as to eliminate the incompatibility.

According to the scorekeeping model, attention must be given not only to "modal" incompatibility but also to "normative" incompatibility. Again, modal incompatibility refers to states of affairs and properties of *objects* that are incompatible with what others and it presupposes the world as independent of the attitudes of the knowing-and-acting subjects.

Normative incompatibility belongs to discursive practices on the side of the knowing-and-acting subjects. In discursive practice the agent cannot be entitled to incompatible doxastic or practical commitments and if one finds herself in this situation one is obliged to rectify or repair the incompatibility. On the side of the object, it is impossible for it to have *incompatible* properties at the same time; on the side of the subject, it is *impermissible* to have incompatible commitments at the same time. In this sense, we can introduce the metaphysical categorical sortal metaconcept subject whereas it represents the conceptual functional role of *units of account for deontic normative incompatibilities*. In my opinion, we can intend this role as a "social" role because of the fact that we learn how to undertake deontic attitudes in the process of socialization. The possibility of criticizing commitments in order to be able not to acknowledge incompatible commitments is bound to the normative statuses of commitment and entitlement and we ought to grasp the sense of them.

The scorekeeping model describes a system of social practices in which agents performs assertions that express material inferential commitments (Giovagnoli 2018). In the previous section, I considered together with the modal vocabulary also the normative vocabulary both related to the use of ordinary language. Let's see now what are the inferential relations that agents ought to master in order for justifying their claims. Our assertions have a "sense" or are "contentful" by virtue of three dimensions of inferential social practices. To the first dimension belongs the *commitment-preserving* inference that corresponds to the material deductive inference. For example, A is to the west of B then B is to the east of A and the *entitlement preserving* inference that corresponds to inductive inference like if this thermometer is well made then it will indicate the right temperature. This dimension is structured also by *incompatibility* relations: two claims have materially incompatible contents if the commitment to the one precludes the entitlement to the other.

The second dimension concerns the distinction between the *concomitant* and the *communicative* inheritance of deontic statuses. To the concomitant inheritance corresponds the *intrapersonal* use of a claim as a premise. In this case, if a person is committed to a claim is, at the same time, committed to other concomitant claims as consequences. Correspondently, a person entitled to a commitment can be entitled to others by virtue of permissive inferential relations. Moreover, incompatibility relations imply that to undertake a commitment has as its consequence the loss of the entitlement

to concomitant commitments to which one was before entitled. To the communicative inheritance corresponds the interpersonal use of a claim, because to undertake a commitment has as its "social" consequence to entitle others to the "attribution" of that commitment. The third dimension shows the two aspects of the assertion as "endorsed": the first aspect is the "authority" to other assertions and the second aspect dependent to the first is the "responsibility" through which an assertion becomes a "reason" enabling the inheritance of entitlements in social contexts.

The entitlement to a claim can be justified (1) by giving reasons for it, or (2) by referring to the authority of another agent, or (3) by demonstrating the capacity of the agent reliably to respond to environmental stimuli. The scorekeeping model is based on a notion of entitlement that presents a structure of "default" and "challenge". This model is fundamental in order to ground a pragmatic and social model of justification, that requires the participation to the game of giving and asking for reasons. A fundamental consequence of this description is that the deontic attitudes of the interlocutors represent a perspective on the deontic states of the entire community.

We begin with the intercontent/interpersonal case. If, for instance, B asserts "That's blue", B undertakes a doxastic commitment to an object being blue. This commitment ought to be attributed to B by anyone who is in a position to accept or refuse it. The sense of the assertion goes beyond the deontic attitudes of the scorekeepers, because it possesses an inferentially articulated content that is in a relationship with other contents. In this case, if by virtue of B's assertion the deontic attitudes of A change, as A attributes to B the commitment to the claim "That's blue", then A is obliged to attribute to B also the commitment to "That's colored". A recognizes the correctness of that inference when she becomes a scorekeeper and, therefore, consequentially binds q to p. Again, the incompatibility between "That's red" and "That's blue" means that the commitment to the second precludes the entitlement to the first. Then A treats these commitments as incompatible if she is disposed to refuse attributions of entitlement to "That's red" when A attributes the commitment to "That's blue". In the infracontent/interpersonal case, if A thinks that B is entitled (inferentially or not inferentially) to the claim "That's blue", then this can happen because A thinks that C (an agent who listened to the assertion) is entitled to it by testimony.

An interesting point is to see how the inferential and incompatibility relations among contents alter the score of the conversation. First, the scorekeeper A must include "That's blue" in the set of commitments already attributed to B. Second, A must include the commitments to whatever claim which is the consequence of "That's blue" (in commitive-inferential terms) in the set of all the claims already attributed to B. Second, A must include the commitment to whatever claim which is the consequence of "That's blue" (in commitive-inferential terms) in the set of all the claims already attributed to B. This step depends on the available auxiliary hypothesis i relationship with other commitments already attributed to B. These moves determine the closure of the attributions of A to B by virtue of the commitment-preserving inferences: starting from a priori context with a certain score, the closure is given by whatever committive-inferential role A associates with "That's blue" as part of its content. Naturally, the resulting attributions of entitlements must not be affected by material incompatibility.

Incompatibility also limits the entitlements attributed to B. A can attribute entitlements to what ever claim is a consequence in permissive-inferential terms of

commitment to which B was already entitled. It can be, however, the case that B is entitled to "That's blue" because she is a reliable reporter i.e. she correctly applies responsive capacities to environmental stimuli. The correctness of the inference depends here on A's commitment, namely on the circumstances under which the deontic status was acquired (these conditions must correspond to the ones in which B is a reliable reporter of the content of "That's blue"). Moreover, A can attribute the entitlement also by inheritance: reliability of another interlocutor who made the assertion in a prior stage comes into play.

6 Conclusion

The pragmatic model I sketched could represent a valid perspective for social epistemology by virtue of its "relational" perspective. It rests on social evidence that derive from semantic relations among material-inferential commitments and entitlements and pragmatic attitudes expressed by a net of basic speech acts. The structure represents a view of knowledge as projected by the discursive practices of an entire community of language users. Moreover, it is a dynamic model as social practices are always exposed to the risk of dissent. In this context, social practices entail the dimension of challenge, i.e. the case in which the speaker challenges the interlocutor to justify and eventually to repudiate his/her commitment. Even in the case in which an agent acquires the entitlement to act by deferral i.e. by indicating a testimonial path whereby entitlements to act can be inherited, the query and the challenge assume the function of fostering the reflection among the participants.

Acknowledgements. I would like to thank Lorenzo Magnani and the participants to the MBR18-Spain for their fruitful comments. I am grateful to Matthieu Fontaine and the reviewers for their careful work and patience.

References

Brandom R (1994) Making it explicit. Cambridge University Press, Cambridge
Brandom R (1995) Knowledge and the social articulation of the space of reasons. Philos Phenomenol Res 55:895–908
Brandom R (2000) Articulating reasons. Harvard University Press, Cambridge, pp 71–72
Brandom R (2008) Between saying and doing. Oxford University Press, Oxford
Brandom R (2012) How analytic philosophy has failed cognitive science. In: Brandom R (ed) Reason in philosophy: animating ideas. Harvard University Press, Cambridge, cfr, pp 26–29
Dummett M (1973) Frege: philosophy of language. Duckworth, London
Dummett M (1991) The logical basis of metaphysics. Harvard University Press, Cambridge
Giovagnoli R (2001) On normative pragmatics: a comparison between brandom and habermas. Teorema 20(3):51–68
Giovagnoli R (2004) Razionalità espressiva. Scorekeeping: inferenzialismo, pratiche sociali e autonomia, Mimesis, Milano
Giovagnoli R (2007) Autonomy: a matter of content. FUP, Florence

Giovagnoli R, Seddone G (2009) Autonomia e intersoggettività. Aracne, Roma
Giovagnoli R (2012) Why the fregean square of opposition matters for epistemology. In: Beziau JY, Dale J (eds) Around and beyond the square of opposition, Birkhauser, Springer, Basel
Giovagnoli R (2013) Representation, analytic pragmatism and AI. In: Dodig-Crnkovic G, Giovagnoli R (eds) Computing nature. Springer, Heidelberg, pp 161–170
Giovagnoli R (2017a) The Relevance of language for the problem of representation. In: Dodig-Crnkovic G, Giovagnoli R (eds) Representation and reality: humans, other living beings and intelligent machines. Springer, Basel
Giovagnoli R (2017b) Introduzione all'epistemologia sociale. LUP, Vatican City
Giovagnoli R (2018) From single to relational scoreboards. In: Beeziau JY, Costa-Leite A, D'Ottaviano JML (eds) Aftermath of the logical paradise, Colecao CLE, Brazil, vol 81, pp 433–448
Goldman A (2015) Social Epistemology, Stanford Enciclopedia of Philosophy
Hardwig J (1985) Epistemic dependence. J Philos 82:1985
Hardwig J (1991) The role of trust to knowledge. J Philos 88:693–708
Kusch M (2002) Knowledge by agreement. Oxford University Press, Oxford
Lewis D (1983) Scorekeeping in a language game. Philosophical Papers. Oxford University Press, New York
Murzi J, Steinberger F (2017) Inferentialism. In: Hale B, Wright C, Miller A (eds) A companion to the philosophy of language. Wiley, New Jersey
Rosemberg JF (1997) Brandom's making it explicit: a first encounter. Philos Phenomenol Res LVII:179–187
Sellars W (1956) Counterfactuals, dispositions and the causal modalities. In: Feigl H, Scriven M, Maxwell G (eds) Minnesota studies in the philosophy of science, vol. II, University of Minnesota Press, Minneapolis
Weiss B, Wanderer J (eds) (2010) Reading brandom. Routledge, Abingdon
Welbourne M (1993) The community of knowledge. Aldershot, Gregg Revivals



Counterfactual Thinking in Cooperation Dynamics

Luís Moniz Pereira[1(✉)] and Francisco C. Santos[2,3]

[1] NOVA-LINCS and Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisbon, Portugal
lmp@fct.unl.pt
[2] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[3] ATP-Group, IST-Taguspark, Porto Salvo, Portugal

Abstract. Counterfactual Thinking is a human cognitive ability studied in a wide variety of domains. It captures the process of reasoning about a past event that did not occur, namely what would have happened had this event occurred, or, otherwise, to reason about an event that did occur but what would ensue had it not. Given the wide cognitive empowerment of counterfactual reasoning in the human individual, the question arises of how the presence of individuals with this capability may improve cooperation in populations of self-regarding individuals. Here we propose a mathematical model, grounded on Evolutionary Game Theory, to examine the population dynamics emerging from the interplay between counterfactual thinking and social learning (i.e., individuals that learn from the actions and success of others) whenever the individuals in the population face a collective dilemma. Our results suggest that counterfactual reasoning fosters coordination in collective action problems occurring in large populations, and has a limited impact on cooperation dilemmas in which coordination is not required. Moreover, we show that a small prevalence of individuals resorting to counterfactual thinking is enough to nudge an entire population towards highly cooperative standards.

Keywords: Counterfactuals · Cooperation · Evolutionary Game Theory

1 Introduction

Counterfactual Thinking (CT) is a human cognitive ability studied in a wide variety of domains, namely Psychology, Causality, Justice, Morality, Political History, Literature, Philosophy, Logic, and AI [1–8]. CT captures the process of reasoning about a past event that did not occur, namely what would have happened had the event occurred, which may take into account what we know today. CT is also used to reason about an event that did occur, concerning what would have followed if it had not; or if another event might have happened in its place. An example situation: Lightning hits a forest and a devastating forest fire breaks out. The forest was dry after a long hot summer and many acres were destroyed. A counterfactual thought is: If only there had not been lightning, then the forest fire would not have occurred.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 69–82, 2019.
https://doi.org/10.1007/978-3-030-32722-4_5

Given the wide cognitive empowerment of CT in the human individual, the question arises of how the presence of individuals with CT-enabled strategies affects the evolution of cooperation in a population comprising individuals of diverse interaction strategies. The natural locus to examine this issue is Evolutionary Game Theory (EGT) [9], given the amount of extant knowledge concerning different types of games, strategies, and techniques for the evolutionary characterization of such populations of individual players. EGT represents a dynamical and population-based counterpart of classical game theory. Having originated in evolutionary biology, EGT also holds a great potential in the realm of social sciences, given that human decision-making is often concerned within large populations and networks of self-regarding individuals.

The framework of EGT has been recently used to identify several of the key principles of mechanisms underlying the evolution of cooperation in living systems [9–11]. Most of these principles have been studied within the framework of two-person dilemmas. In this context, the Prisoner's Dilemma (PD) metaphor is possibly the most ubiquitously known game of cooperation. The dilemma is far more general than the associated story, and can be easily illustrated as follows. Two individuals have to decide simultaneously to offer (to *Cooperate*) or not (to *Defect*) a benefit b to the other at a personal cost $c < b$. From a game theoretical point of view a rational individual in a PD is always better off by not cooperating (defecting), irrespectively of the choice of the opponent, while in real life one often observes the opposite, to a significant extent. This apparent mismatch between theory and empirical results can be understood if one assumes an individual preference to cooperate with relatives, direct reciprocity, reputation, social structure, direct positive and negative incentives, among other sorts of community enforcing mechanisms (for an overview, see, e.g., [10, 11]).

Despite the popularity of the PD, other game metaphors can be used to unveil the mysteries of cooperation. Each game defines a different metaphor, the relative success of a player and attending strategy, and the ensuing behavioural dynamics. Moreover, while 2-person games represent a convenient way to formalize a pairwise cooperation, many real-life situations are associated with dilemmas grounded on decisions made by groups of more than 2 agents. Indeed, from group hunting, to modern collective endeavours such as *Wikipedia* and open source projects, or global collective dilemmas such as the management of common pool resources or the mitigation of the dangerous effects of climate change, general N-person problems are recurrent in biological and social settings.

The prototypical example of this situation is the N-person Prisoner's dilemma, also known as Public Goods Game. Here, N individuals decide to contribute (or not) to a *public good*. The sum of all contributions is invested, and the returns of the investment shared equally among all group members, irrespectively of who contributed. Often, in these cases, free riding allows one to enjoy the public good at no cost, to the extent that others continue to contribute. If all players adopt the same reasoning, we are led to the *tragedy of the commons* [12], characterizing the situation in which everyone defects, with the making of cooperation a mirage. Below we will return to the details associated with the formalization of such dilemmas.

Importantly, depending on the game and associated strategies, individuals may revise their strategies in different ways. The common assumption of classic game theory is that players are rational, and that the Nash Equilibrium constitutes a

reasonable prediction of what self-regarding rational agents adopt [13]. Often, however, players have limited cognitive skills or resort to simpler heuristics to revise their choices. Evolutionary game theory (EGT) [14] offers an answer to this situation, adopting a population description of game interactions in which individuals resort to social learning and imitation. In EGT, the accumulated returns of the game are associated with the fitness of an individual, such that the most successful individuals would reproduce more often, and their strategies spread in the populations. Interestingly, such evolutionary framework is mathematically equivalent to social learning, where individuals revise their strategy looking at and imitating the success and actions of others that show more fit than they [9]. As a result, strategies that do well spread in the population.

Yet, contrary to social learning, more sophisticated agents (such as humans) might instead imagine how a better outcome could have turned out, if they would have decided differently, and thence self-learn by revising their strategy. This is where Counterfactual Thinking (CT) comes in. In this chapter, we propose a simple mathematical model to study the impact on cooperation of having a population of agents resorting to such counterfactual kind of reasoning, when compared with a population of just social learners. Specifically, in this chapter, we pose three main questions:

1. How can we formalize counterfactual behavioural revision in large populations (taking cooperation dynamics as an application case study)?
2. Will cooperation emerge in collective dilemmas if, instead of evolutionary dynamics and social learning, individuals revise their choices through counterfactual thinking?
3. What is the impact on the overall levels of cooperation of having a fraction of counterfactual thinkers in a population of social learners? Does cooperation benefit from such diversity in learning methods?

To answer these questions, we develop a new population dynamics model based on evolutionary games, which allows for a direct comparison between the behavioural dynamics created by individuals who revise their behaviours through social learning and through counterfactual thinking. We consider individuals who interact in a public goods problem in which a threshold of players, less than the total group size, is necessary to produce benefits, with increasing contributions leading to increasing returns. This setup is common to many social dilemmas occurring in Nature and societies [15–17], combining a N-player interaction setting with non-linear returns. We show that such counterfactual players can have a profound impact on the levels of cooperation. Moreover, we show that just a small prevalence of counterfactual thinking within a population is sufficient to lead to higher overall levels cooperation when compared with a population made up solely of social learners.

To our knowledge, this is the first time that counterfactual thinking is considered in the context of the evolution of cooperation in populations of agents, by employing evolutionary games with both counterfactual thinking and social learning to that effect. Nevertheless, other works have made use of counterfactuals in the multi-agent context. Foerster et al. [18] addressed cases where counterfactual thinking consists in imagining changes in the rules of the game, an interesting problem not addressed in this chapter. Peysakhovich et al. [19] rely on the availability and use of a centralized critic

that estimates counterfactual advantages for the multi-agent system by reinforcement learning policies. While interesting from a engineering perspective, it differs significantly from our work, in that we have no central critic, no utility function to be maximized, nor the aim of conjuring a policy that optimizes given utility function. To the contrary, in our approach the population evolves, leading to emergent behaviours, by relying on social learning and counterfactual thinking given at the start to some of the agents. Hence the two approaches deal with distinct problem settings and are not comparable. A similar approach is used by Colby et al. [20] where there is a collective utility function to be maximized. Finally, in Ref. [7], it is adopted a modelling framework that neither considers a population nor multi-agent cooperation, but individual counterfactual thinking about another agent's intention. It does this by counterfactually imagining whether another agent's goal would still be achieved if certain harmful side effects of its actions had not occurred. If so, those harmful side effects were not essential for the other achieving its goal, and hence its actions were not immoral. Otherwise, they were indeed immoral because the harmful side effects were intended because necessary to achieve the goal.

This chapter is organized as follows. In Sect. 2 we detail the principles underlying our counterfactual thinkers and the N-person collective dilemma used to illustrate the idea. In Sect. 3 we introduce the mathematical formalism associated with counterfactual and social learning dynamics. Importantly, this formalism is independent of the game chosen. In Sect. 4 we show the results for the impact on cooperation dynamics of counterfactual reasoning when compared with social learning, and discuss the influence of counterfactual thinkers in hybrid population of both social learners and counterfactual agents. Section 5 provides a discussion on the results obtained, also drawing some suggestions for future works.

2 Counterfactual Thinking and Evolutionary Games

Counterfactual thinking (CT) can be exercised after knowing one's resulting payoff following a single playing step with a co-player. It employs the counterfactual thought: *Had I played differently, would I have obtained a better payoff than I did?* This information can be easily obtained by consulting the game's payoff matrix, assuming the co-player would have made the same play, that is, other things being equal. In the positive case, the CT player will learn to next adopt the alternative play strategy.

A more sophisticated CT would search for a counterfactual play that improves not just one's payoff, but one that also contemplates the co-player not being worse off, for fear the co-player will react negatively to one's opportunistic change of strategy. More sophisticated still, the new alternative strategy may be searched for taking into account that the co-player also possesses CT ability. And the co-player might too employ a Theory of Mind (ToM)-like CT up to some level. We examine here only the non-sophisticated case, a model for simple (egotistic) CT.

In Evolutionary Game Theory (EGT), a frequent standard form of learning is so-called Social Learning (SL). It basically consists in switching one's strategy by imitating the strategy of a more successful individual in the population, compared to one's success. CT instead can be envisaged as a form of strategy update learning akin to

debugging, in the sense that: *if my actual play move was not conducive to a good accumulated payoff, then, after having known the co-player's move, I can imagine how I would have done better had I made a different strategy choice*.

When compared with SL, this type of reasoning is likely to have a minor impact in games of cooperation with a single Nash equilibrium (or a single evolutionary stable strategy, in the context of EGT) such as the Prisoner's dilemma or the Public Goods game mentioned above, where defection-dominance prevails. However, as we illustrate below, counterfactual thinking has the potential to have a strong impact in games of coordination, characterized by multiple Nash Equilibria: CT will allow for a meta-reasoning on which equilibria provide higher returns. This is particularly relevant since conflicts often described as public goods problems, or as a Prisoner's Dilemma game, can be also interpreted as coordination problems, depending on how the actual game parameters are ranked and assessed [15].

As an example, consider the two-person Stag-Hunt (SH) game [15]. Players A and B decide to go hunt a stag, which needs to be done together for maximum possibility of success. But each might defect and go by himself hunt a hare instead, which is more assured, because independent of what the other does, though less rewarding. The dilemma is that each is not assured the other will in fact go hunt the stag in cooperation and is tempted by the option of playing it safe by going hunt hare. Concretely, one may assume that hunting hare has a payoff of 3, no matter what the other does; hunting stag with another has a payoff of 4; and hunting stag alone has a payoff of 0. Hence the two-person stag hunt expectance payoff matrix:

	B stag	**B** hare
A stag	4 4	0 3
A hare	3 0	3 3

A simple analysis of this payoff table would tell us that one should always adopt the choice of our opponent, i.e., to coordinate our actions, despite the fact that we may end in a sub-optimal *equilibria* (both going for hare).

The nature of this dilemma is generalizable to an N-player situation where a group of N is required to hunt stag [17]. Let us then consider a group of N individuals, who can be either cooperators (C) or defectors (D); the k Cs in N contribute a cost c to the public good, whereas Ds refuse to do so. The accumulated contribution is multiplied by an enhancement factor F, and the ensuing result equally distributed among all individuals of the group, irrespective of whether they contributed or not. The requirement of coordination is introduced by noticing that often we find situations where a minimum number of Cs is required within a group to create any sort of collective benefit [17, 21]. From group hunting [17] and the rise and sustainability of Human organizations [22], to collective action to mitigate the effects of dangerous climate change [23], examples abound where a minimum number of contributions is required before any public good is produced.

Following [17], this can be modelled through the addition of a coordination threshold M, leading to the following straight forward payoff functions for the Ds and

Cs—Cs contribute a cost c to the common pool and Ds refuse to do so—where j stands for the number of contributing k:

$$\begin{aligned} P_D(j) &= H(j-M)jFc/N \\ P_C(j) &= P_D(j) - c \end{aligned} \tag{1}$$

respectively. Here, H represents the Heaviside step function, taking the value of $H(x) = 1$ if $x \geq 0$, and $H(\mathrm{x}) = 0$ otherwise. Above the coordination threshold M, the accumulated contribution ($j.c$) is increased by an enhancement factor F, and the total amount is equally shared among all the N individuals of the group. This game is commonly referred to as the N-person Stag-Hunt game [17]. For $M = 1$, one recovers the classical N-person Prisoner's dilemma and the Public Goods game alluded to in the introduction.

Here we will consider a population of agents facing this N-person Stag-hunt dilemma, revising their preferences through social learning and through counterfactual thinking. The following section provides the details of how the success of an agent is computed and how agents revise their strategies in each particular case. The mathematical details are given for the sake of completeness, but a detailed understanding of the equations is not required to follow the insights of the model. For more information on evolutionary game theory and evolutionary dynamics in finite population, we refer the interested reader to reference [9].

3 Population Dynamics with Social Learning and Counterfactual Thinking

Let us consider finite population of Z interacting agents where all agents are equally likely to interact with each other. In this case, the success (or fitness) of an agent results from the average payoff obtained from randomly sampling groups of size $N < Z$. As a result, all individuals adopting a given strategy will share the same fitness. One can compute the average fitness of each individual through numerical simulations, averaging over the payoff received in a large number of groups randomly sampled from the population. Equivalently, one can also compute analytically the average fitness of a strategy assuming a random sampling of groups, and averaging over the returns obtained in each group configuration[1].

[1] Formally, one can write this process as an average over a hyper-geometric sampling in a population of size Z and k cooperators. This gives the probability of an agent to interact with N-1 other players, where, among those, j are cooperators. In this case, the average fitness f_D and f_C of Ds and Cs, respectively, in a population with k Cs, is given by [17] $f_D(k) = \binom{Z-1}{N-1}^{-1} \sum_{j=0}^{N-1} \binom{k}{j} \binom{Z-k-1}{N-j-1} P_D(j)$ and $f_C(k) = \binom{Z-1}{N-1}^{-1} \sum_{j=0}^{N-1} \binom{k-1}{j} \binom{Z-k}{N-j-1} P_C(j+1)$, where P_C and P_D are given by Eq. (1).

Using this framework as a baseline model for the interactions among agents, let us now detail how population evolution proceeds under social learning (SL) and under counterfactual thinking (CT). Our Z interacting agents can either resort to SL or CT to revise their behaviours. If an agent i resorts to SL, i will imitate a randomly chosen individual j, with a probability p that augments with the increase in fitness difference between j and i, given by f_i and f_j respectively. This probability can take many forms. We adopt the ubiquitous standard Fermi distribution to define the probability p for SL

$$p_{SL} = \left[1 + e^{-\beta_{SL}\left[f_j - f_i\right]}\right]^{-1} \tag{2}$$

in which β_{SL} expresses the unavoidable noise associated with errors in the imitation process [24]. Hence, successful individuals might be imitated with some probability, and the associated strategy will spread in the population.

Given the above assumptions, it is easy to write down the probability to change the number k of Cs (by ± 1 at each time step) in a population of $Z - k$ Ds in the context of social learning. The number of Cs will increase if an individual D has the chance to meet a role model with the strategy C, which occurs with a probability $\frac{Z-k}{Z}\frac{k}{Z-1}$. Imitation will then effectively occur with a probability $\left[1 + e^{-\beta_{SL}[f_C - f_D]}\right]^{-1}$ (see Eq. 2), where f_C and f_D are the fitness of a C and of a D, respectively. Similarly, to decrease the number of Cs, one needs a C meeting a random role model D, which occurs with a probability $\frac{k}{Z}\frac{Z-k-1}{Z-1}$; imitation will then occur with a probability given by Eq. 2. Altogether, one may consequently write the probability that one more or one less individual (T^+ and T^- in Fig. 1) adopts a cooperative strategy through social learning as[2]

$$T^{\pm}_{SL}(k) = \frac{k}{Z}\frac{Z-k}{Z}\left[1 + e^{\beta_{SL}[f_B(k) - f_A(k)]}\right]^{-1}. \tag{3}$$

Differently, agents that resort to CT assess alternatives to their present returns, had they used an alternative play choice *contrary* to what actually took place. Agents imagine how the outcome could have worked if their decision (or strategy, in this case) would have been different. In its simplest form, this can be modelled as an incipient form of myopic best response rule [13] at the population level, taking into account the fitness of the agent in a configuration that did not, but could have occurred. In the case of CT, an individual i adopting a strategy A will switch to B with a probability

$$p_{CT} = \left[1 + e^{-\beta_{CT}\left[f_i^B - f_i^A\right]}\right]^{-1} \tag{4}$$

that increases with the fitness difference between the fitness the agent would have had if it had played B (f_i^B), and the fitness the agent actually got by playing A (f_i^A). This can easily be computed considering the fitness players of C or players of D would have had in a population having an additional cooperator, or an extra defector, depending on the alternative strategy chosen by the individual revising its strategy. As before, one may

[2] For simplicity we assume that the population is large enough such that $Z \approx Z - 1$.

write down the probability to change the number k of Cs by plus or minus 1. To increase the number of Cs, one would need to select a D—which occurs with a probability $(Z-k)/k$—and this D counterfactually decides to switch to C with the probability given by Eq. (4):

$$T_{CT}^{+}(k) = \frac{Z-k}{Z}\left[1+e^{-\beta_{CT}\left[f_C(k+1)-f_D(k)\right]}\right]^{-1} \tag{5}$$

Similarly, the probability to decrease the number of Cs by one through counterfactual thinking would be given by

$$T_{CT}^{-}(k) = \frac{k}{Z}\left[1+e^{-\beta_{CT}\left[f_D(k-1)-f_C(k)\right]}\right]^{-1} \tag{6}$$

This expression is slightly simpler than the transition probabilities of SL due to the fact that, in this case, the reasoning is purely individual, and does not depend on the existence and fitness of meeting individuals adopting a different strategy.

Importantly, CT assumes that agents get access to the returns under different actions. This is, of course, a strong assumption that precludes the use of CT in all levels of complexity and evolutionary settings. Nonetheless, we assume that this feature is shadowed by some sort of error through the parameter β_{CT}, which, once again, expresses the noise associated with guessing the fitness values.

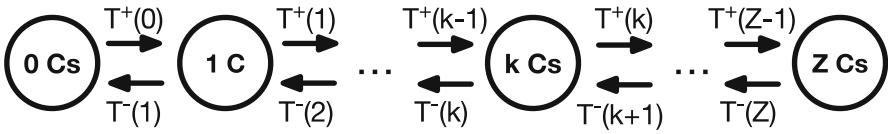


Fig. 1. Social dynamics as a Markov chain. Representation of the transitions among the $Z + 1$ available states, in a population of size Z, and two strategies (C and D). Both social learning $T_{SL}^{\pm}(k)$ and counterfactual thinking $T_{CT}^{\pm}(k)$, where k stands for the number of cooperators (Cs), create a one-dimensional birth and death stochastic process of this type. One may also inquire about the stationary distribution s_k for each case, which represent the expected fraction of time a population spends in each given state, in a large timespan.

We further assume that, with probability μ, individuals may switch to a randomly chosen strategy, freely exploring the space of possible behaviours. Thus, with probability μ, there occurs a mutation and an individual adopts a random strategy, without resorting to any of the above fitness dependent heuristics. With probability $(1 - \mu)$ we have either social learning or counterfactual reasoning learning. As a result, for both SL and CT transition probabilities, we get modified transition probabilities given by

$$T^{+}_{SL/CT}(k,\mu) = (1-\mu)T^{+}_{SL/CT}(k) + \mu(Z-k)/Z \tag{7}$$

for the probability to increase from k to $k + 1$ Cs and

$$T^{-}_{SL/CT}(k,\mu) = (1-\mu)T^{-}_{SL/CT}(k) + \mu k/Z \tag{8}$$

for the probability to decrease to $k - 1$ (see Fig. 1).

These transition probabilities can be used to assess the most probable direction of evolution in SL and CT. This is given by a learning gradient (often called gradient of selection [17, 23] in the case of SL), expressed by

$$G_{SL}(k) = T^{+}_{SL}(k,\mu) - T^{-}_{SL}(k,\mu) \tag{9}$$

and

$$G_{CT}(k) = T^{+}_{CT}(k,\mu) - T^{-}_{CT}(k,\mu) \tag{10}$$

respectively. When $G_{SL}(k) > 0$ and $G_{CT}(k) > 0$ ($G_{SL}(k) < 0$ and $G_{CT}(k) < 0$), time evolution is likely to act to increase (decrease) the number of Cs. In other words, for a given number k of Cs, the sign of $G(k)$ offers the most likely direction of evolution.

This mathematical framework allows one to obtain the fraction of time that the population spends in each configuration after a long time has elapsed. To do so, it is important to note that the above transition probabilities define a stochastic process in which the probability of each event depends only on the current state of the population. In other words, we are facing a Markov process, whose states are given by the number of cooperators $k \in \{0, \ldots, Z\}$. The transitions among all $Z + 1$ states can be seen as a transition matrix Λ_{ij} such that $\Lambda_{k,k\pm1} = T^{\pm}_{SL/CT}(k,\mu)$ and $\Lambda_{k,k} = 1 - \Lambda_{k,k-1} - \Lambda_{k,k+1}$. The average time the system spends in each state k is given by the so-called *stationary distribution* s_k, which is obtained from the eigenvector corresponding to the eigenvalue 1 of the transition matrix Λ [25].

Finally, we can also use the stationary distribution to define a global *cooperation index*

$$\langle C \rangle = \sum_k k s_k \tag{11}$$

which gives the number of Cs across states (k), weighted by the time the system spends in each state, i.e., by the stationary distribution s_k.

4 A Comparison of Social Learning and Counterfactual Prompted Evolutions, and an Analysis of Their Interplay

In Fig. 2a, we illustrate the behavioural dynamics both under CT and SL for the same parameters of the N-person Stag-hunt game. For each fraction of co-operators (Cs), if the gradient G (for both SL or CT) is positive (negative), then it is likely the fraction of Cs will increase (decrease). As shown, in both cases, the dynamics is characterized by two basins of attraction and two interior fixed points[3]: one unstable (also known as a coordination point), and a stable co-existence state between Cs and Ds. To achieve stable levels of cooperation (in a co-existence state), individuals must coordinate to be able to reach the cooperative basin of attraction on the right-hand side of the plot, a common feature in many non-linear public goods dilemmas [17, 23]. Figure 2a also shows that CT allows for the creation of new playing strategies, absent before in the population, since new strategies can appear spontaneously based on individual reasoning. By doing so, CT interestingly leads to different results if compared to SL. In this particular scenario, it is evident how CT may facilitate coordination of action, as individuals can reason on the sub-optimal outcome associated with non-reaching the coordination threshold, and individually react to that.

In Fig. 2b, we show the stationary distribution of the Markov chain associated with the transition probabilities indicated above, showing how cooperation can benefit from CT. The stationary distribution characterizes the prevalence in time of each fraction of co-operators (k/Z). In this particular configuration, it is shown how in SL (black line), the population spends most of the time in low values for the fraction of co-operators. Whenever CT is allowed, cooperation is maintained most of the time. This emerges from the new position of the unstable fixed point shown in Fig. 2a. We further confirmed (not shown) the equivalence of CT and SL prompted evolutions, in the absence of coordination constraints (i.e., when $M = 1$). In this case, we would have a defection dominance dilemma with a single attractor at 100% of defectors. Thus, in this regime, CT will have a marginal impact. Nonetheless, as long as the N-person game includes the need for coordination, translated in the existence of (at least) two basins of attraction and an internal unstable fixed point, CT may have the positive impact shown in Fig. 2. In this particular case of the N-person Stag-Hunt dilemma, as shown by Pacheco et al. [17], the existence of these two basins of attraction depends on the interplay between F, M and the group size, N.

[3] Strictly speaking, by the finite population analogues of the internal fixed points in infinite populations.

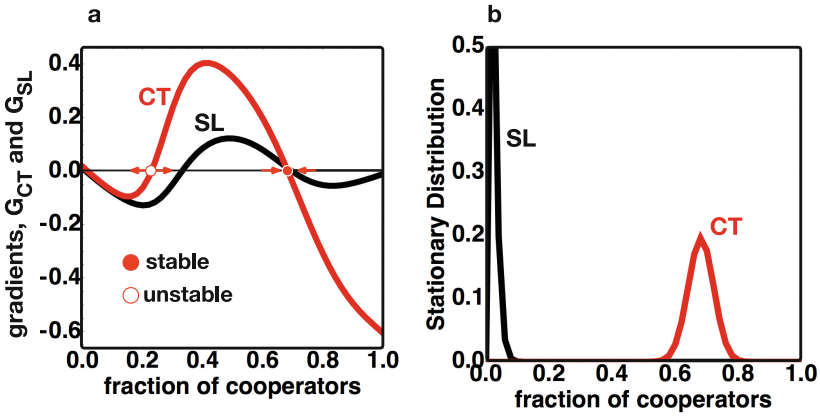


Fig. 2. Left panel: learning gradients for social learning (black line) and counterfactual thinking (red line)—see $G_{SL}(k/Z)$ in see Eq. 9 and $G_{CT}(k/Z)$ in Eq. 10—for the N-person SH game ($Z = 50$, $N = 6$, $F = 5.5$. $M = N/2$, $c = 1.0$, $\mu = 0.01$, $\beta_{SL} = \beta_{CT} = 5.0$). For each fraction of co-operators, if the gradient is positive then it is likely that the number of co-operators will increase; for negative gradient values cooperation is likely do decrease. Empty and full circles represent the finite population analogue of unstable and stable fixed points, respectively. Right panel: Stationary distribution of the Markov processes created by the transition probabilities pictured in the left panel; it characterizes the prevalence in time of each fraction of co-operators in finite populations (see main text). Given the positions of the roots of the gradients and the amplitudes of G_{CT} and G_{SL}, contrary to the SL configuration, in the CT case the population spends most of the time in a co-existence between Cs and Ds.

Until now, individuals can either revise their strategies through social learning or counterfactual reasoning. However, one could also envisage situations where each agent may resort to CT and to SL in different circumstances, a situation prone to occur in Human populations. To encompass such heterogeneity at the level of agents, let us consider a simple model in which agents resort to SL with a probability χ, and to CT with a probability $(1 - \chi)$, leading to a modified learning gradient given by $G(k) = \chi G_{SL}(k) + (1 - \chi)G_{CT}(k)$.

In Fig. 3, we show the impact χ on the average cooperation levels (see Eq. 11) in a N-person Stag-Hunt dilemma in which, in the absence of CT, cooperation is unlikely to persist. Remarkably, our results suggest that a tiny prevalence of individuals resorting to CT is enough to nudge an entire population of social learners towards highly cooperative standards, providing further indications on the robustness of cooperation prompted by counterfactual reasoning.

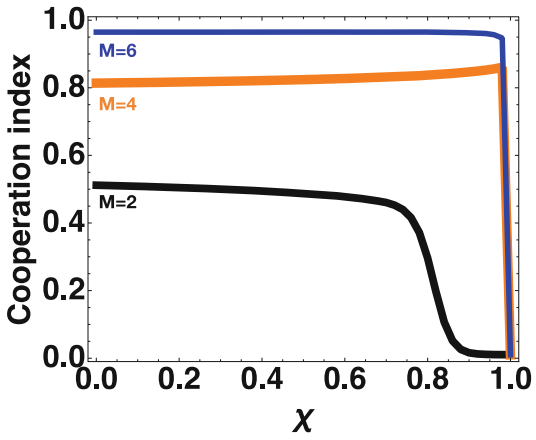


Fig. 3. Overall level of cooperation (cooperation index, see Eq. 11) as a function of the prevalence of individuals resorting to social learning (SL, χ) and counterfactual reasoning (CT, $1 - \chi$). We show that only a relatively small prevalence of counterfactual thinking is required to nudge cooperation in an entire population of self-regarding agents. Other parameters: $Z = 50$, $N = 6$, $F = 5.5.$, $c = 1.0$, $\mu = 0.01$, $\beta_{SL} = \beta_{CT} = 5.0$.

5 Discussion

In this contribution we illustrate how decision-making shaped by counterfactual thinking (CT) is worth studying in the context of large populations of self-regarding agents. We propose a simple form to describe the population dynamics arising from CT and study its impact in behavioural dynamics when individuals face a threshold public goods problem. We show that CT enables the arise of pro-social behaviour in collective dilemmas, even where non or little existed before. We do so in the framework of non-cooperative N-person evolutionary games, showing how CT is able to modify the *equilibria* expected to occur when individuals revise their choices through social learning (SL). Specifically, we show that CT is particularly effective in easing the coordination of actions by displacing the unstable fixed points that characterize this type of dilemmas. In the absence of a clear need to coordinate (e.g., whenever $M = 1$ in the N-person game discussed) CT offers equivalent results to those obtained with SL. Nonetheless, this is an especially gratifying result since many of the mechanisms known to promote cooperation in defection-dominance dilemmas (e.g., Prisoner's Dilemma, Public Goods games, etc.) enlarge the chances of cooperation by transforming the original dilemma into a coordination problem [10]. Thus, CT has the potential to be even more effective when applied in combination with other known mechanisms of cooperation, such as conditional strategies based on past actions, commitments, signalling, emotional reactions, or reputations based dynamics [16, 26–28]. Moreover, it is worth pointing out that the NPSH dilemma adopted here as a particular case study—which combines co-existence and coordination dynamics (see Fig. 2a)—also represents the dynamics that emerge from most N-person threshold games, and from standard Public Goods dilemmas in the presence of group reciprocity

[29], quorum sensing [16], and adaptive social networks [30], which further highlights the generality of the insights provided.

We also analyse the impact of having a mixed population of CT and SL, since it is unlikely that individuals would resort to a single heuristic for strategy revision. We show that, even when agents seldom resort to CT, highly cooperative standards are achieved. This result may have various interesting implications, if heterogeneous populations are considered. For instance, we can envision a near future made of hybrid societies comprising humans and machines [31–33]. In such scenarios, it is not only important to understand how human behaviour changes in the presence of artificial entities, but also to understand which properties should be included in artificial agents capable of leveraging cooperation in such hybrid collectives [31]. Our results suggest that a small fraction of artificial CT agents in a population of Humans social learners can decisively influence the dynamics of cooperation towards a cooperative state. These insights should be confirmed through a two-population ecological model where SLs influence CTs (and vice-versa), but also by including CTs that have access to a lengthier record of plays, rather than just the last one, or learn from past actions, creating a time-dependence that may be assessed through numerical computer simulations. Work along these lines is in progress.

Acknowledgements. We are grateful to The Anh Han and Tom Lenaerts for comments. We are also grateful to the anonymous reviewers for their improvement recommendations This work was supported by FCT-Portugal/MEC, grants NOVA-LINCS UID/CEC/04516/2013, INESC-ID UID/CEC/50021/2013, PTDC/EEI-SII/5081/2014, and PTDC/MAT/STA/3358/2014.

References

1. Mandel DR, Hilton DJ, Catellani P (2007) The psychology of counterfactual thinking. Routledge
2. Birke D, Butter M, Köppe T (2011) Counterfactual thinking-counterfactual writing. Walter de Gruyter
3. Roese NJ, Olson J (2014) What might have been: the social psychology of counterfactual thinking. Psychology Press
4. Wohl V (2014) Probabilities, hypotheticals, and counterfactuals in ancient greek thought. Cambridge University Press
5. Dietz E-A, Hölldobler S, Pereira LM (2015) On conditionals. In: Gottlob G, et al (eds) Global conference on artificial intelligence (GCAI 2015). Citeseer, pp 79–92
6. Nickerson R (2015) Conditional reasoning: the unruly syntactics, semantics, thematics, and pragmatics of "if". Oxford University Press
7. Pereira LM, Saptawijaya A (2017) Counterfactuals, logic programming and agent morality. In: Urbaniak R, Payette P (eds)Applications of formal philosophy: the road less travelled, Logic, Argumentation & Reasoning series. Springer, pp 25–54
8. Byrne R (2019) Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In: international joint conference on AI (IJCAI 2019)
9. Sigmund K (2010) The calculus of selfishness. Princeton University Press
10. Nowak MA (2006) Five rules for the evolution of cooperation. Science 314:1560–1563
11. Rand DG, Nowak MA (2013) Human cooperation. Trends Cogn Sci 17:413–425
12. Hardin G (1968) The tragedy of the commons. Science 162:1243

13. Fudenberg D, Tirole J (1991) Game theory. MIT Press
14. Hofbauer J, Sigmund K (1998) Evolutionary games and population dynamics. Cambridge University Press
15. Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press
16. Pacheco JM, Vasconcelos VV, Santos FC, Skyrms B (2015) Co-evolutionary dynamics of collective action with signaling for a quorum. PLoS Comput Biol 11:e1004101
17. Pacheco JM, Santos FC, Souza MO, Skyrms B (2009) Evolutionary dynamics of collective action in N-person stag hunt dilemmas. Proc R Soc Lond B 276:315–321
18. Foerster JN, et al. (2018) Counterfactual multi-agent policy gradients. In: Thirty-second AAAI conference on artificial intelligence 2018, pp 2974–2982
19. Peysakhovich A, Kroer C and Lerer A (2019) Robust multi-agent counterfactual prediction. arXiv preprint arXiv:1904.02235
20. Colby MK, Kharaghani S, HolmesParker C, Tumer K (2015) Counterfactual exploration for improving multiagent learning. In: Proceedings of the 2015 international conference on autonomous agents and multiagent systems: international foundation for autonomous agents and multiagent systems, pp 171–179
21. Souza MO, Pacheco JM, Santos FC (2009) Evolution of cooperation under N-person snowdrift games. J Theor Biol 260:581–588
22. Bryant J (1994) Problems of coordination in economic activity. Springer, pp 207–225
23. Santos FC, Pacheco JM (2011) Risk of collective failure provides an escape from the tragedy of the commons. Proc Natl Acad Sci USA 108:10421–10425
24. Traulsen A, Nowak MA, Pacheco JM (2006) Stochastic dynamics of invasion and fixation. Phys Rev E 74:011909
25. Karlin S, Taylor HMA (1975) A first course in stochastic processes. Academic
26. Han TA, Pereira LM, Santos FC (2011) Intention recognition promotes the emergence of cooperation. Adapt Behav 19:264–279
27. Han TA, Pereira LM, Santos FC (2012) The emergence of commitments and cooperation. In: Proceedings of AAMAS 2012, IFAAMS. pp 559–566
28. Santos FP, Santos FC, Pacheco JM (2018) Social norm complexity and past reputations in the evolution of cooperation. Nature 555:242
29. Van Segbroeck S, Pacheco JM, Lenaerts T, Santos FC (2012) Emergence of fairness in repeated group interactions. Phys Rev Lett 108:158104
30. Moreira JA, Pacheco JM, Santos FC (2013) Evolution of collective action in adaptive social structures. Sci Rep 3:1521
31. Paiva A, Santos FP, Santos FC (2018) Engineering pro-sociality with autonomous agents. In: Thirty-second AAAI conference on artificial intelligence 2018, pp 7994–7999
32. Santos FP, Pacheco JM, Paiva A, Santos FC (2019) Evolution of collective fairness in hybrid populations of humans and agents. In: Thirty-third AAAI conference on artificial intelligence 2019, vol 33, pp 6146–6153
33. Pereira LM, Saptawijaya A (2016) Programming machine ethics. Springer



Modeling Morality

Walter Veit(✉)

University of Bristol, Bristol, UK
wrwveit@gmail.com

Abstract. Unlike any other field, the science of morality has drawn attention from an extraordinarily diverse set of disciplines. An interdisciplinary research program has formed in which economists, biologists, neuroscientists, psychologists, and even philosophers have been eager to provide answers to puzzling questions raised by the existence of human morality. Models and simulations, for a variety of reasons, have played various important roles in this endeavor. Their use, however, has sometimes been deemed as useless, trivial and inadequate. The role of models in the science of morality has been vastly underappreciated. This omission shall be remedied here, offering a much more positive picture on the contributions modelers made to our understanding of morality.

Keywords: Morality · Evolution · Replicator dynamics · Moral dynamics · Models · Evolutionary game theory · ESS · Explanation

1 Introduction

Since Axelrod's (1984) famous work *The Evolution of Cooperation*[1], economists, biologists, neuroscientists, psychologists, and even philosophers have been eager to provide answers to the puzzling question of why humans are not the selfish creatures natural selection seems to demand.[2] The list of major contributions is vast. Of particular importance is Skyrms' pioneering use of evolutionary game theory (abbreviated as EGT) and the replicator dynamics in his books *The Evolution of the Social Contract* (1996) and *The Stag Hunt and the Evolution of Social Structure* (2004). Further important book-length contributions on the evolution of morality are offered by Wilson (1975); Binmore (1994, 1998, 2005); de Waal (1996, 2006); Sober and Wilson (1998); Joyce (2006); Alexander (2007); Nowak and Highfield (2011); Bowles and Gintis (2011); Boehm (2012), and most recently Churchland (2019).

The efforts of these and many other authors have led to the formation of an interdisciplinary research program with the explicit aim to explain and understand human morality, taking the first steps towards a genuine science of morality. Let us call this research program the *Explaining Morality Program* (**EMP**). For a variety of reasons, models, such as those provided by Skyrms, have played a very important role in this endeavor, the most illustrative reason being the simple fact that behavior does not fossilize. Models and simulations *alone*, however, have been doubted by many to

[1] Based on an earlier co-authored paper with Axelrod and Hamilton (1981) of the same name.

[2] See Dawkins' (1976) book *The Selfish Gene* for an elegant illustration of the problem.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 83–102, 2019.
https://doi.org/10.1007/978-3-030-32722-4_6

provide much of an explanation when it comes to human morality. The work of modelers in the **EMP** has been underappreciated for a number of reasons that can roughly be grouped together in virtue of the following three concerns: (i) the complexity of the phenomenon, (ii) the lack of empirical support, and perhaps the most threatening criticism being (iii) the supposedly non-reducible normative dimension morality embodies.[3] In this paper, I shall argue that this underappreciation is a mistake.[4]

Though interdisciplinarity has played a crucial role in the advancement of our moral understanding; it has led to an underappreciation of the role and contribution that highly abstract and idealized models have played. Many responses to the modeling work within the **EMP** are characterized by eager attempts to draw lines in the sand, i.e. determine prescriptive norms that would limit the justified use or misuse of such models.[5] These criticisms range from sophisticated ones, perhaps the most convincing one offered in Levy (2011) to rather naïve criticisms such as those offered in Arnold (2008). The latter goes so far as to label such models as useless, trivial and inadequate. In a harsh review of Arnold (2008); Zollman (2009) criticized Arnold's arguments against the use of models, deeming them unconvincing and exceedingly ambitious. Modelers may very well be tempted to attack Arnold as a straw-man and conclude that the criticism of models for the evolution can easily be debunked. However, such an approach would ignore the more sophisticated arguments that have been offered.

For the purposes of debunking the strongest arguments against such models, any mention of Arnold's (2008) criticism hardly deserves mention here. Arnold is by no means alone, however. His mistakes illustrate a shared pattern that can be found, though in a much weaker form, across the literature. It is a sort of a priori skepticism and perhaps dislike among philosophers and more experimentally oriented scientists, about the role of models in science. This skepticism is one I hope to at least partially dispose of here.[6] I shall demonstrate that models for the evolution of morality are neither too simplistic, nor do they lack empirical data to provide us with genuine explanatory insights. Recent advances in the philosophical literature on models, especially on model pluralism and the role of multiple models, should allow us to recognize not only such often exaggerated limitations but also the *strengths* of models

[3] See Rosenberg and Linquist (2005); Northcott and Alexandrova (2015); Nagel (2012) respectively as examples for each.

[4] The EMP has faced similar criticism itself, relating not only to mathematical models but towards scientific explanations of morality at large. My goal here is only to provide a defence of the models used in this research program. I suggest, however, that if my attempt succeeds the entire EMP justifies its status as a genuine science of morality. Nevertheless, see FritzPatrick (2016) for a recent overview of EMP critics and defenders alike.

[5] See D'Arms (1996, 2000); D'Arms et al. (1998); Rosenberg and Linquist (2005); Nagel (2012); Northcott and Alexandrova (2015); Arnold (2008); Kitcher (1999); Levy (2011, 2018).

[6] Godfrey-Smith (2006), for instance, diagnoses a general distrust among philosophers in respect to "resemblance relations [of models] because they are seen as vague, context-sensitive, and slippery" (p. 733). Similarly Sugden (2009) has argued that models work by a form of induction "however problematic [that] may be for professional logicians" (p. 19).

in the **EMP**.[7] The latter of which have often been underappreciated, while the former have been overstated. This omission shall be remedied here.

In order to demonstrate a number of conceptual mistakes made in the literature, I shall largely draw on Alexander's (2007) book, *The Structural Evolution of Morality*, offering perhaps the most extensive modeling treatment on the evolution of morality. Building on previous work by his former supervisor, Skyrms (1996, 2004), Alexander analyzes a large scope of exceedingly complex and arguably more realistic models, in order to illuminate the requirements and potential threats to the emergence and spread of moral behavior.[8] He concludes that morality can be explained by a combination of evolutionary game theory (abbreviated as **EGT**), together with a theory of bounded rationality and research in psychology. In doing so, he attempts to answer two distinct questions: (i) how could something as altruistic as human morality emerge and (ii) how did it persist against the threat of cheaters?

For the purposes of this paper, Alexander's (2007) book serves as a highly attractive case study for two reasons. Firstly, Alexander's contribution relies solely on highly abstract and idealized models of precisely the form often criticized as too simplistic to provide us with genuine insights for phenomena as complex as human morality. Secondly, while economists, psychologists, biologists, neuroscientists, and even political scientists have provided substantial contributions to the **EMP**, philosophers have offered distinct and extremely valuable insights by drawing conceptual distinctions.[9] As both a modeler and philosopher, Alexander treads very carefully only suggesting possible insights, his book may provide. At times, he even underestimates his own scientific contribution, arguing that it does not tell us much, if anything, without the supplementation of much more empirical data. One may regard such humility as a virtue, but at times, even an unbiased reader may get the impression that Alexander himself sees his contribution as superfluous. However, all of this humility seems to be thrown overboard in the very end of his book, where he discusses and suggests implications of the **EMP** for our understanding of morality itself, vindicating the 'objective status' of morality. Nevertheless, despite giving in too much to the criticisms of the program, Alexander avoids several pitfalls that might obscure our understanding of the epistemic contribution such models can provide. Here I shall shed a much more positive light on the role of models in the **EMP**, or as it sometimes referred to as: the study of *moral dynamics*.[10]

The structure of this paper corresponds roughly to the three concerns raised against the role of models in the **EMP** illustrated above. Firstly, in Sect. 2, I discuss Alexander's contribution and explore the most important question within the literature, i.e. why model morality? In Sect. 3, I respond to concerns regarding their empirical

[7] See Knuuttila (2011); Muldoon (2007); Wimsatt (2007); Weisberg (2007a, 2013); Ylikoski and Aydinonat (2014); Lisciandra (2017); Aydinonat (2018); Grüne-Yanoff and Marchionni (2018).

[8] To some extent one may treat his contribution as an extended robustness analysis of Skyrms' prior work. This would not do justice to Alexander's contribution, however.

[9] Joyce (2006) book *The Evolution of Morality* offers perhaps the most valuable contribution in this regard.

[10] See Hegselmann (2009).

adequacy, before finally, in Sect. 4, I cast doubt on the possibility of vindicating the objective status of moral norms *via* the **EMP** and conclude the discussion.

2 Why Model Morality?

Evolutionary explanations of morality have been of scientific interest, since at least Darwin. Proto-Darwinian explanations, however, have been around for a long time. As Hegselmann (2009) points out, the **EMP** has a long scientific tradition going back as far as ancient Greece. Protagoras, in one of Plato's dialogues, provides perhaps the first scientific explanation of morality as a set of norms and enforcement agencies being an invention of humanity to escape a Hobbesian state of nature.[11] Couched in terms of a myth, we may treat this as a mere just-so story. Much later, David Hume came astonishingly close to providing a Darwinian explanation of morality himself.[12] Hegselmann and Will (2013) determine four key components to Hume's proto Darwinian account: a pre-societal human nature with confined generosity, the invention of artificial values to be reinforced and internalized through approval and disapproval of others, division of labour reaping the benefits of cooperation and trust and the "invention of central authorities that monitor, enforce, and eventually punish behaviour" (p. 186) already much more sophisticated but still similar to the myth of Prometheus and Epimetheus told by Protagoras. These accounts, a mere story and myth in the case of Protagoras and in the case of Hume an informal suggestion of a how-possibly explanation leave much to be desired, but they were, nevertheless, the best explanations available at the time. Luckily it didn't take two millennia for the next advancement. Joyce (2006) in his book *The Evolution of Morality*, argues that "less than a century after Hume's death, Darwin delivered the means for pushing the inquiry into human morality further" (p. 228) filling out a gap Hume could only describe as *nature*. Charles Darwin, of course, himself suggested that the origin of morality can be explained with his theory:

> It may be well first to premise that I do not wish to maintain that any strictly social animal, if its intellectual faculties were to become as active and as highly developed as in man, would acquire exactly the same moral sense as ours. In the same manner as various animals have some sense of beauty, though they admire widely different objects, so they might have a sense of right and wrong, though led by it to follow widely different lines of conduct. If, for instance, to take an extreme case, men were reared under precisely the same condition as hive-bees, there can hardly be a doubt that our unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering. Nevertheless, the bee, or any other social animal, would gain in our supposed case, as it appears to me, some feeling of right or wrong, or a conscience. (1879, p. 67)

This, of course, is still, 'just' a how-possibly explanation or as critics like to call them a just-so story. It should be clear that from Protagoras over Hume to Darwin,

[11] Plato (1961). Protagoras. In: Hamilton E, Huntington C (eds) The collected dialogues of Plato. Princeton University Press, Princeton.

[12] Hume (1998). An enquiry concerning the principles of morals (ed by Beauchamp TL). Oxford University Press, Oxford.

significant improvements in the explanation of morality have been made with more and more gaps being closed. Explanations come in degrees, and this research program is providing better and better explanations, and perhaps the best for the moment. Unfortunately, it took a while until informal evolutionary explanations resting on the good for the species were replaced with formal **EGT** models showing that the origin of moral behavior is not much of a mystery after all. Skyrms (1996) was the first to apply evolutionary game theory to unpack Hume's account in a formal manner, with others following in the creation of new models and simulations.[13] These sets of models strengthen our confidence that morality could have evolved in a way envisioned by Hume and Darwin, providing considerable explanatory power, even though empirical work has, hitherto, been largely left out of the picture.[14] Even the work of moral philosophers in this Humean research program has been very empirical and guided by science trying to unpack the idea of morality being a mere artefact (see Mackie 1977; Joyce 2001, 2006) being a case in point for the division of labor between philosophers, modelers and empirical researchers. Modelers such as Skyrms simply continue an old philosophical school of thought with the modern tools of science, a move that ought to be encouraged.

Before engaging in a more detailed analysis of our case study, i.e. the models Alexander (2007) provides, I shall take on his last chapter titled "Philosophical reflections" where he explores the philosophical implications of his models. Though the appearance of moral behavior in Alexander's models is rather robust and remains stable even in the face of defectors, more he argues needs to be said in order to draw inferences about human morality. Quoting Philip Kitcher, Alexander (2007, p. 267) highlights a general problem for evolutionary explanations of morality:

> [I]t's important to demonstrate that the forms of behaviour that accord with our sense of justice and morality can originate and be maintained under natural selection. Yet we should also be aware that the demonstration doesn't necessarily account for the superstructure of concepts and principles in terms of which we appraise those forms of behaviour (Kitcher 1999).

In response, Alexander introduces a distinction between "thinly" and "thickly" conforming to morality. Though an individual's action may conform thinly with morality, e.g. fair sharing, the individual may fail "to hold sufficiently many of the beliefs, intentions, preferences, and desires to warrant the application of the term ['] moral[‘] to his or her action" (2007, p. 268). In contrast, thickly conforming to morality satisfies sufficiently many of these conditions. If someone acts 'morally' out of purely selfish reasons, we may not want to call such behavior moral, e.g. someone giving to the poor in order to improve their reputation. Akin to Kant, it is the distinction between behavior in compliance with morality or acting out of the right, i.e. moral reasons.

When evolutionary game theory models are used to simulate the emergence and persistence of moral behavior we only observe the "frequencies and distribution of strategies and, perhaps, other relevant properties" (Alexander 2007, p. 270). What is lacking here is the role of psychology, perhaps even neuroscience, in the production of such moral behavior. Even if we allow for very complex strategies, such as those

[13] See Alexander (2007); Hegselmann and Will (2010, 2013).

[14] See also O'Connor (forthcoming).

submitted to Axelrod's (1984) computer tournament, they still allow for a purely behavioral interpretation.

This problem lies at the core of attempts to model the evolution of morality. Critics argue that a complete explanation for the evolution of morality requires an understanding of the internal psychological mechanisms that produce such moral behavior. Alexander concedes to this criticism, suggesting to enrich these models with "non-strategic, psychological elements" (p. 273). He grants that **EGT** alone is not sufficient for an evolutionary explanation of morality, but that "together with experimental psychology and recent work in the theory of bounded rationality […] some of the structure and content of our moral theories" can be explained "by working in tandem" (p. 274). This position, of course, is a much weaker one than to claim that **EGT** *alone* could provide genuine insights into the origins of morality.

But even if, as I suggest, **EGT** might be sufficient to explain much of our moral behavior, Alexander aims at more. First of all, as Kitcher suggests, evolutionary game theory enables the important identification of behavior that maximizes long-run expected utility or fitness. A second step then is required to explain the motivational structures which are "actually *producing* this behavior in boundedly rational individuals" (2007, p. 275). Here Alexander identifies two mechanisms. First, the moral sentiments bringing about motivation to act, and secondly, moral theories instructing us "how to act once we have been motivated to do so" (275). Therefore, Alexander argues, it is precisely because we are boundedly rational that the "outcome produced by acting in accordance with moral theory are such that they tend to maximize our expected utility over the lifetime of the individual" (p. 275). Rationality requires us to rely on heuristics, and these luckily, according to Alexander, are often moral heuristics such as a fair split and cooperation.[15]

In analyzing the influence of moral heuristics on our thinking, Alexander, based on a distinction by Sadrieh et al. (2001), discusses three separate roles that moral heuristics play in our thinking. Firstly, moral heuristics limit our set of options, e.g. not even considering to poison the dog of our neighbor, even though his barking may disturbs one's sleep. Secondly, moral heuristics guide our information search, i.e. what we need to consider before making a judgement. Thirdly, but closely related to the second point, moral heuristics "tell us when to terminate an information search" (p. 277). When we find out that someone killed a human infant for fun, it is sufficient for a moral judgement regardless of any additional information. Dennett (1996a, b) has defended a similar position on moral judgements, calling them *conversation-stoppers* for otherwise costly debates. Relatedly, Alexander makes the rather contentious claim, that though we use moral reasoning, moral theories have their form precisely because they track "long-run expected utility" (p. 278). The key to the evolution of morality he argues "lies in the fact that we all face repeated interpersonal decision problems – of many types – in socially structured environments" (p. 278), hence the *structural* evolution of morality.

As "the science of morality is only in its infancy" (p. 281) there must remain some unanswered questions in our current explanation, however and I agree here with

[15] See Veit et al. (forthcoming) for an analysis of the 'Rationale of Rationalization'.

Alexander, this "is no reason why we should not make the attempt" (p. 282). Akin to primates, evolution equipped us with "emotions and other cognitive machinery" (p. 284), in order to solve interdependent decision problems, such as those arising in the prisoner's dilemma, the stag hunt, and the divide the cake game. Rosenberg makes a similar argument and extends it to love, as the "solution to a strategic interaction problem" (2011, p. 3). Analogously, the mere fact that love is an evolved response does not have to undermine our conviction that the feelings and intentions associated with it are not genuine or worthy of pursuit. The same may hold for morality. Emotions and our cognitive machinery is the raw material evolution had to use in order to solve more complex problems humans were increasingly facing, e.g. trust and the introduction of property rights.[16] With the evolution of language, this arms race in human evolution could only gain speed. We do not know yet which of our moral attitudes are hard-wired and which are culturally acquired, but that is obviously no reason not to ask the question. As we shall see, many of the EGT models used by Alexander, Skyrms, and others allow for both a cultural and biological interpretation. As cultural evolution operates at a much higher speed; however, many modelers such as Alexander (2007) give them a cultural interpretation.

Establishing the motivation for his models, Alexander henceforth, turns to the evidential support for evolutionary explanations of morality, in order to turn them into more than 'just-so stories', i.e. evolutionary explanations without empirical evidence. Evolutionary explanations are often faced with the criticism of providing nothing more than 'just-so stories', i.e. historical accounts without any empirical evidence in their favour. For Charles Darwin, it was very important to collect plentiful evidence for his theory of natural selection and biologists to this day continue to accumulate corroborating evidence. However, when biologists try to explain the occurrence of a certain behavior or a phenotype in general, they often start by hypothesizing how the trait could be adaptive. This research program is often criticized as a sort of *Panglossian adaptationism*, i.e. assuming the adaptiveness of a trait without further evidence.[17] Though Alexander only considers two experiments, (see Yaari and Bar-Hillel 1984; Binmore et al. 1993), they are highly suggestive that our conception of fairness is somewhat flexible and strongly correlates with the outcomes our own group receives. Though a philosophical review of the vast literature on moral experiments should be undertaken, it is beyond the scope and purpose of this paper, which is merely concerned with attempts to model morality.[18] Nevertheless, the just-so story critique has evolved into a term of abuse used against all kinds of model-based explanations. In this paper, I shall attempt to argue against this commonplace treatment and highlight the wealth and diversity models can provide in the **EMP**.

Though brought up as a game theorist, Alexander recognizes the weakness in the assumptions of standard rational choice theory. In order to avoid charges of unrealisticness, Alexander's models for the evolution of moral behavior make no strong

[16] Here the often drawn distinction between biological and psychological altruism plays an important role.

[17] See Gould and Lewontin 1979 for their famous critique of adaptationism.

[18] See Kagel and Roth (1998) for an overview of such studies in experimental economics.

rationality assumptions; rather he uses models of bounded rationality combined with evolutionary game theory to account for the evolution of morality. Sugden anticipated as much in a paper on the evolutionary turn in game theory stating that the "theory of human behaviour that underlies the evolutionary approach is fundamentally different from that which is used in classical game theory" (2001, p. 127), with far less contestable rationality assumptions, though similar in their mathematical formulation. In short: Alexander treats bounded rationality theory as descriptively superior to standard rational choice theory.

However, with the threat of only providing so-called *just-so stories*, evolutionary explanations, in general, are often dismissed by pointing to the multiplicity of evolutionary accounts we could give for the appearance of a phenomenon. These objections, however, miss the mark when they supposed to show that evolution plays no part in explaining morality. Alexander's former supervisor, Brian Skyrms, himself working on the evolution of social norms makes this criticism of just-so story charges explicit: "Why have norms of fairness not been eliminated by the process of evolution? […] How then could norms of fairness, of the kind observed in the ultimatum game, have evolved?" (1996, p. 28). In this section, I argue that such criticism is highlighting something important that Sugden (2000; 2009) tries to capture in his work on model-based explanations. Though very similar arguments have been made by Giere (1988, 1999), Godfrey-Smith (2006), Weisberg (2007b, 2013), and Levy (2011), Sugden's work serves as an elegant illustration of Alexander's aims for at least two reasons: (i) Sugden's account is partially motivated by evolutionary game theory models used in both economics and biology, and (ii) his 'credible world' terminology maps neatly onto the justifications, goals, and inferences Alexander is drawing himself.

Models, Sugden (2000, 2009) argues are parallel worlds, artificially created, which can be used to draw inductive inferences to the real world. At least he argues, such is the practice in economics and biology. In both of these fields, phenomena are complex and can be multiply realized by different mechanisms. This is why, Sugden argues, we need induction to bridge the gap between the model world and the real world, even though he grants that this may seem unappealing to some philosophers. A model here, in virtue of its idealizations, is a sort of fictional entity that enables us to draw inductive inferences about the real world via similarity relations to the 'model world'. Hence, Sugden argues, modelers aim to create 'credible worlds' that we *could imagine being real*. It is not truth *per se* that is aimed for, but rather a sort of credibility that is deemed able to tell us something about the real world we live in. To do so modelers, are required to provide us with relevant similarities between what is happening in the model and what could be going on in the real world, perhaps requiring a sort of elaborative story or narrative linking the two. In the following, I argue that Alexander's contribution to the **EMP** consists in the construction of such 'credible worlds' from which we can draw inductive or abductive inferences to the real world.[19]

[19] I treat abduction, i.e. inference to the best explanation, here similar to Sugden (2009) as a form of induction. Others do not share this view, instead arguing that eliminative induction is a form of IBE e.g. Aydinonat (2007, 2008). However, I have no bone to pick in this debate. What conception one holds does not impact the validity of the arguments presented here.

Analyzing the phenomena cooperation, trust, fairness and retribution, Alexander (2007) conducts his project by exploring different and increasingly complex models in which he wants to explore the evolution of morality.[20] He goes on to employ five models, i.e. replicator dynamics, lattice models, small-world networks, bounded-degree networks and dynamic networks each introducing more and more elaborate forms of population structure back into the picture and increasing the realism of his models. He analyses four different dimensions of morality, i.e. cooperation, trust, fairness and retribution amounting to a set of twenty models, each having their robustness tested in several iterations. Each of these models alone seems to tell us very little about the real world. Taken together, however, this extensive set of robust models supports Alexander's assertion that population structure plays a very important role in the evolution of morality.

First, he starts with a simple model used in evolutionary biology and increasingly the social sciences, i.e. the replicator dynamics. As already alluded to, **EGT** allows for both biological and cultural interpretations explaining the interdisciplinary interest in **EGT**. While the biological form of these models treats replication as (biological) inheritance, replication has to be interpreted as some form learning or imitation in a cultural setting. Replicator dynamics (RD) are an attempt to model the relative changes of strategies in a population. Again, these can be either instantiated biologically or culturally. Strategies with higher fitness than the population average prosper and increase their share in the population, while those with lower fitness are driven to extinction. RD in the biological setting are thus an attempt to model the dynamics of reproduction and natural selection. The following is the continuous replicator dynamics equation:

$$\frac{dx_i}{dt} = [u(i,x) - u(x,x)] * x_i \quad \text{(Weibull, 1995, p. 72)} \tag{1}$$

In each round individual strategies, i increase their share within a population linear to their success u(i, x) compared to the average fitness u(x, x) in the population. Just as the evolutionary stable strategy (ESS) familiar from earlier evolutionary game theory models, RD assume infinite population size or at least infinite divisibility and random interaction. These idealizations allow us to analyze the frequency-dependent success of different strategies, whether they are biologically or culturally transmitted. Though he intends his project to model the cultural evolution of morality, he grants that replicator dynamics leave it open whether the strategies are genetically or culturally transmitted. Let us consider Alexander's first example and the most-analysed game in game theory: the *Prisoner's Dilemma*.

Table 1. The payoff matrix for the Prisoner's Dilemma

	Cooperate	Defect
Cooperate (Lie Low)	R	S
Defect (Anticipate)	T	P

[20] See Gelfert (2016) for a recent discussion of the various exploratory functions of models.

In the *Prisoner's Dilemma,* there is only one NE, i.e. Defect, Defect. This famous game can be traced back to Hobbes (1651), who argued that a powerful leader is required to escape the state of nature, i.e. collective defection. In fact, his name is mentioned over twenty times in Alexander's book, pointing to the long tradition of the **EMP**. In Table 1, T is "temptation", i.e. the value tempting defection, R is the "reward" of joint cooperation, P is "punishment" as both receive a lower payoffs then they would have gotten if both had cooperated, and S is the "sucker's" payoff where a co-operator is exploited (2007, p. 55). The payoffs are ordered as follows $T > R > P > S$ with the additional condition that $T + S/2$ is smaller than R. The ESS here coincides with the strict Nash Equilibrium (**NE**)[21] predicting mutual defection. Using replicator dynamics to model the evolutionary trajectory shows that co-operators are quickly driven to extinction, with defectors taking over the population.

As his book is called *The Structural Evolution of Morality*, Alexander is aware that human societies are more complex and that we need to account for the social structure of society in order make these models more *credible*. In fact, when population structure is introduced, and interactions are no longer entirely random, making it possible for co-operators to group together, cooperation can persist and evolve. Therefore he moves on to explore agent-based models, i.e. lattice models, small-world networks, bounded-degree networks and dynamic networks, where agents can choose with whom to interact. Increasing the complexity in his models serves then two purposes: on the one hand the goal is (i) to ensure robustness, i.e. the stability of the outcomes in the model under changes in the model and on the other hand (ii) to increase the credibility of the model, i.e. the likeliness of it telling us the true story about the evolution of morality. As Sugden says: "what we need in addition is some confidence that the production model is likely to do the job for which it has been designed – that it is likely to explain real-world phenomena" (2000, p. 11), and this is Alexander's stronger aim: the pre-vision of a *how actual* explanation.[22] Let me therefore, now tackle these two purposes in succession.

Looking at robustness first, Alexander claims that the results in his models are sufficiently robust to suggest that moral behavior can emerge and remain stable in a population of boundedly rational agents. I agree with Kuorikiski and Lehtinen (2009) that robustness analysis is somewhat implicit in Sugden's account of inductive inference[23], this, however, should be interpreted as a continuous inference from increasingly similar model worlds to their real-world counterpart. Robustness analysis and inductive inference are closely related and overlap in important respects, but Sugden is justified

[21] The Nash Equilibrium introduced by John Nash, is the central, in fact, most important solution concept in Game Theory. The concept picks out a combination of strategies, i.e. one for each 'player' in the game, in which none of the players has an incentive to unilaterally deviate from his chosen strategy, while the strategies others have chosen remain fixed. In the Prisoner's Dilemma this classically leads to only one unique solution, i.e. mutual defection. Morality quickly suggests itself as an evolved social solution to such inefficient equilibria.

[22] I use how actual explanations in a modal sense, i.e. as a subset of how possibly explanations.

[23] One may even treat robustness analysis as a necessary component of model-based science itself. Sometimes it is used in a very narrow sense, at other times quite broadly. See Lisciandra (2017) for a recent overview, but also Woodward (2006)

in making a distinction on the grounds of their different epistemic properties. Robustness analysis increases internal validity for the model world, while this internal validity is a prerequisite for establishing external validity in the real world. When slightly changed alterations of the model are seen as the target themselves, this distinction breaks down. I take Levy (2011) subtle criticism on the modeling literature on the evolution of morality, to target the tendency of not sufficiently distinguishing between these two distinct ways in which validity can be increased. Modelers such as Skyrms, Levy argues, take their models to establish external validity, when really only internal validity has been vindicated. I will say more about Levy's criticism in the next section on the empirical adequacy of these models.

When models are used to learn about human morality, Sugden (2009), Cartwright (2009) and others are correct in arguing that the purpose of models is to learn something about a real-world target system. Francesco Guala, argues that it is "necessary to investigate empirically which factors among those that may be causally relevant for the result are likely to be instantiated in the real world but are absent from the experiment (or vice versa)" (2005, p. 157). This procedure of establishing external validity not only applies to inductive inference from the artificial experimental world to the real world but also to inductive inference from the artificial model world to the real world. In both cases, we want to draw inferences from highly idealized and abstract mechanisms to a causal mechanism operating in the real world. As several authors have recently pointed out, there are more important similarities than relevant differences between models and experiments, which makes it difficult to justify drawing any hard boundaries between the two.[24]

In order to gain confidence that Alexander's 'story' provides us with the actual explanation of human morality, requires more, especially evidence from psychology and neuroscience, in order to learn about the causal mechanism behind moral behavior. Even though our models are robust, this robustness in itself only tells us something about the evolution of moral behavior in the model, i.e. unless relevant similarities obtain between the real and the model world. Sugden argues that "a transition has to be made from a particular hypothesis, which has been shown to be true in the model world, to a general hypothesis, which we can expect to be true in the real world too" (2000, p. 19), i.e. inductive inference. Sugden explicates three such inductive schemata, explanation, prediction and abduction. For the purpose of this paper, only his explanation schema is relevant:

E1 – in the model world, R is caused by F.
E2 – F operates in the real world.
E3 – R occurs in the real world.
Therefore, there is a reason to believe:
E4 – in the real world, R is caused by F. (2000, p. 12)

The phenomena R in question is the emergence and stability of moral behavior. Though Alexander explicitly wants to explain more, i.e. the emergence and stability of *morality*, we shall first consider whether these models can explain 'moral' behavior.

[24] See Mäki (2005) and Parke (2014)

What we need to establish in order to make justified inductive inferences, i.e. extrapolation, from the model world to the real world are relevant similarities. While the relevant set of causal factors in the model, i.e. cultural evolution, do operate in the real world this may be an unavoidable feature of generalized theories. When Sugden speaks of a model's *credibility* he is not talking about their literal truth, but *truth-likeness*, a description of "how the world could be" (2000, p. 24), a credible counterfactual world. For a model world to achieve this kind of credibility, it needs to cohere with the causal processes we know to be operating in the real world. The agents postulated in our model need to be in a relevant sense like real agents in our world. By using evolutionary game theory and bounded rationality models, Alexander intends to trump the standard rational choice theory models in virtue of credibility. Drawing inferences from his models, therefore, he argues is at least inductively more justified than standard game theory explanations for the evolution of morality. This concession, however, is unlikely to convince many critics of rational choice theory.

However, if standard rational choice models are justified, his models should be justified by extension in virtue of their enhanced credibility. If the standard models fail to achieve credibility or rather external validity, we need some further argument, to see how Alexander's models are explanatory. When Alexander moves from the simple replicator dynamics to lattice models, he is continuing his quest for a more credible model world. Here we drop the unrealistic assumption of random interaction in an infinitely sized population for a one-dimensional lattice in which everyone has two neighbors to interact with. Secondly, Alexander analyzes how different learning rules change the strategy dynamics in his models, all of which are rather simple but perhaps better capture the actual strategy changes in human agents. As the assumption of only interacting with two neighbors is highly unrealistic in itself, Alexander moves to small-world networks where some agents have an additional interaction possibility by being connected over a bridge. Further increasing credibility, Alexander moves to bounded-degree networks where every agent has a certain number of connections i between k (min) and k(max). Here connections need not be neighbors and are randomly assigned, creating networks that look fairly similar to interaction networks in the real world. However, humans obviously do not choose with whom to interact entirely at random. When we encounter someone who cheats in a cooperative endeavor we will try to avoid them and interact with someone else next time. Alexander draws on Skyrms and Pemantle (2000) model of social network formation in order to model changing interaction frequencies. Without going into the specifics and intricacies of each of these models, they illustrate an important point: Alexander's book follows the modeling strategy of first ensuring robustness and internal validity, before moving on to more credible model worlds that gain complexity and inferential power. The latter approach must, of course, be closely related to empirical research into morality, most importantly, perhaps moral psychology.

Robert Sugden's account of modeling is justified in virtue of being a naturalistic, pragmatic account of the actual scientific modeling practice. Models are successful in explaining but do so by induction. Therefore, we should accept induction as a valid principle in the modelers toolkit, "however problematic [that] may be for professional logicians" (2009, p. 19). In a research paper on the evolutionary turn in game theory, Sugden writes:

> Evolutionary game theory is still in its infancy. A genuinely evolutionary approach to economic explanation has an enormous amount to offer; biology really is a much better role model for economics than is physics. I just hope that economists will come to see the need to emulate the empirical research methods of biology and not just its mathematical techniques. (2001, p. 128)

Alexander doesn't fall into this trap, for he only sees his mathematical models as a subset of the necessary steps towards a genuine explanation for the evolution of morality. This is where empirical evidence needs to be accumulated, and studies conducted, analyzing development psychology with respect to social norms. Alexander's work, however, guides the way for such empirical research and theory testing to commence, in a field that is still nebulous and wide. How to move from robust EGT models to the real world will be explored in the next section.

3 Empirical Adequacy

The most sophisticated criticism of attempts to model the evolution of morality has recently been offered by Levy (2011). Rather than denying the importance of models, Levy argues that there are two distinct modes of inquiry in which indirect modeling can be used to study otherwise complex phenomena. Levy argues that the work of Skyrms, Alexander, and others is characterized by 'internal' progress, achieved within the model, rather than 'target-oriented' progress where we learn more directly about the target system itself. While Levy does not go as far as to argue that this strategy is pure conceptual exploration[25], he suggests that it is "more conceptual in spirit" aimed at understanding the initial model itself (2011, p. 186). Target-oriented modeling, Levy argues, progresses by "incrementally adding causal information" primarily guided by considerations of empirical adequacy (p. 186). In contrast, models for the evolution of morality explore the "subtleties of a constructed set-up" with empirical adequacy only playing a minor role (p. 186).

Similarly, Sugden suggests that a "model cannot prove useful unless someone uses it, and whoever that person is, he or she will have to bridge the gap between model world and real world" (2009, p. 26). Though Alexander downplays the role of his models by saying that they alone cannot account for much, he suggests that jointly with theories of bounded rationality and research in psychology and economics we can get closer to the actual explanation of how morality evolved. This is the main motivation behind Sugden's (2000, 2009) credible world account of modeling. If models were only about conceptual exploration and providing theorems, any mention of the real world and the relationship of the model to it would be *nothing but* telling a story to sell one's model. Levy (2011) suggests that this is what might be happening in the models provided by Skyrms and Alexander. However, as I shall argue, their modeling strategy of the evolution of moral norms explicitly acknowledges relevant real world factors and successively tries to increase the credibility of his models. In the following, I shall argue that **EGT** models can inform empirical research and vice-versa.

[25] See Hausman (1992).

Rosenberg and Linquist (2005) wrote a paper on evolutionary game theory models for cooperation and how to test them empirically, which will prove highly useful for this section. They argue that we can use archaeology, anthropology, primatology and even gene-sequencing to support **EGT** models. Supporting Alexander, they argue that human cooperation is too sophisticated, conditionalized and domain-general for it to be a genetically hardwired trait. What Alexander does not consider is potential gene-culture co-evolution. But I take him to make a deliberately weaker claim, that even if it turns out that human morality is entirely cultural, his models will be useful. If we find empirical evidence for 'hard-wired' behavior than all the better for an evolutionary explanation of morality. However, in order to have a credible **EGT** model Rosenberg and Linquist argue already requires a lot of substantial assumptions, for example, emotion, reliable memories, a theory of (other) mind(s), language and imitation learning. Experiments in economics and psychology, often done in the form of games, can inform us about how humans act, and how a change in conditions changes human behavior. Such empirical work can then help the modeler to not only increase the credibility of his models but also to eliminate those models would tell a completely misguided story of the evolution of (human) morality.[26]

Rosenberg and Linquist provide the popular example of the big-game hunt hypothesis as an explanation for why humans started cooperating. They point out that the empirical data suggests that big-game hunter was an inferior strategy in comparison to gathering, not even granting the payoffs specified in a stag-hunt game. Though the big-game hunt hypothesis tells a nice story of why humans started to cooperate, we should treat it as even less than a 'just so story' because the evidence suggests that it is most likely false. As an alternative, they propose cooperative child-caring which interestingly also fits the mathematical description of a stag-hunt game. In light of empirical evidence, they argue that modelers should try to alter their stag-hunt models for trust by thinking about the potential payoffs of cooperative child-care rather than the payoffs of cooperative hunting.

Such is the nature of this enterprise: both modeling and empirical research can inform each other in a variety of ways. Though many models will be discarded, this leaves us with a much narrower set of how possible explanations and gets us closer to the actual one. They close their paper by stating that is "not for philosophers to speculate how this research once commenced will eventuate" (p. 156), but it is nevertheless necessary to bring in line the theoretical work done in evolutionary game theory with the empirical data from various field.[27] Otherwise, **EGT** models are nothing but conceptual exploration, and as Sugden points out, modelers should and can aim for more. Nevertheless, Hegselmann (2009) suggests caution against the hope that the "huge gap between macroscopic models of moral dynamics and the known variety of microscopic processes that seem to generate certain assumed overall effects" (p. 5689) can be bridged in the near future, if at all. Criticism directed against

[26] As Zollman (2009) points out in his review of Arnold 2008, models have directly inspired experimental work on the evolution of morality, e.g. Wedekind and Milinki (2000); Seinen and Schram (2006).

[27] Similar arguments have recently been raised against the use of game theoretic tools to explain the evolution of multicellularity. See Veit (2019a).

Alexander, stating that he failed to provide a complete explanation is not nearly as effective when a complete explanation is not reachable anyway. While we might consider Alexander's work as only one of the first steps in getting closer to a complete explanation of human morality, it remains an important step nonetheless. Though the empirical data is weak, or rather because of it, it is so important to combine the research results from different fields. Models, far from being a mere add-on to this research program, appear to be a necessary and integral part of the **EMP**, transforming it more and more into a science in its own right. Let me now conclude this discussion with the controversial debate whether such models have any impact on the *moral* status of morality.

4 Implications for the *Moral* Status of Morality

Alexander concludes that "evolutionary game theory, coupled with the theory of bounded rationality and recent experimental work bridging the gap between psychology and economics, provides what appears to be a radical restructuring of the foundations of moral theory" showing that the content of "moral theories are real and binding" (2007, p. 291), though their content is highly dependent on us and the structure of society. Alexander (2007), rather than providing an evolutionary debunking argument for morality, claims to provide an 'objective but relative' basis for morality in so far as he shows that the principles of morality are in the best long-term interest of everyone, a claim that may seem just as radical. This distinction is important as it is often treated anonymously; however, as I shall argue, Alexander goes one step too far when he treats the instrumental justification of morality as an epistemic justification.[28]

Due to the importance of population structure illustrated in his book, Alexander argues morality is necessarily relative to the structure of society. Morality, he argues, is objective but relative. This is an ambitious suggestion, standing in stark contrast to the careful conclusions Alexander has drawn in the rest of his book, and hence deserves closer inspection. Unlike Joyce (2001, 2006); Street (2006) and Sommers and Rosenberg (2003; Rosenberg 2011), who provide evolutionary debunking arguments for the objectivity of morality, Alexander argues that his models, rather than undermining, are able to vindicate morality. And in doing so, he draws explicitly on Hume:

> [M]uch work has to be undertaken in order to unpack Hume's "certain proposition" that "[T] is only from the selfishness and confin'd generosity of men, along with the scanty provision nature has made for his wants, that justice derives its origin."[29] And, as for the origin of justice, so for the rest of morality. (2007, p. 291).

For Humeans in the tradition of Williams (1981) and arguably Hume himself, notions of objectivity were always somewhat odd. In fact, all of the three evolutionary debunkers mentioned above, see themselves as Humeans. They argue that in light of an

[28] I thank Richard Joyce for suggesting this formulation to me.

[29] *A Treatise of Human Nature*, Book III, part II, section II.

evolutionary explanation for the adaptiveness of moral attitudes and behavior, there is neither a need nor should we endorse any 'magical' property that makes morality somehow objective. Nevertheless, Joyce (2001) in his book *The myth of morality* explores the possibility of vindicating the objectivity of morality by linking it to rationality, as perhaps the strongest candidate view to avoid the conclusions of the error theorist (see Mackie 1977). Alexander's argument for the vindication of morality rests on the same motivation: if it can be shown that it is in everyone's interest to act according to morality, morality can be saved.

For this approach to be successful, Joyce (2001) argues we would need to arrive at some sort of categorical imperative that derives from rationality alone, i.e. precisely the route Kant took to save the status of morality from Hume's philosophy. For Humeans, who see reasons as relative to desires, aims and preferences (perhaps also beliefs), this approach must be futile. The moral heuristics will not only be relative to the structure of the society we live in, but also relative to the aims and desires we have, and hence subjective. They would be nothing more than mere heuristics that apply to the majority of the population in the majority of circumstances. Alexander does not see this as a problem; in fact, he sees it as sufficient for grounding morality as something objective, but nevertheless relative.[30]

However, as I see it the previously provided arguments are sufficient for casting doubt on the project of vindicating the objectivity of morality by pointing to a highly relativistic notion of rationality crucially depending on the social structure of society.[31] For even if we grant that this is a sort of objectivity, it is not what humans refer to when talking about the objectivity of morality, nor is it what metaethicists are usually interested in. Error theorists like Mackie, Joyce and Rosenberg readily accept the debunking. Alexander, however, prefers a more subtle version of what we could mean by moral objectivity. His work captures something important: the advice we give our children, the moral norms we teach are likely to be in their long-term best interest. They are useful heuristics that evolved to reap the benefits of cooperation in strategic interaction problems, and as Alexander points out, they are highly contingent on the social structure of society. Levy (2018) suggests that the models explored in the **EMP** could provide us with insights into the desirability of certain institutions and societal norms, merely in virtue of their stability. For meta-ethics, however, the impact of the **EMP** may be severe. Akin to an electromagnetic pulse (**EMP**) the Explaining Morality Program could paralyze much of the traditional work of philosophers working on morality.[32] Hence, it comes with no surprise that many naturalists and philosophers of

[30] Joyce (2001, 2006) explores these issues in more detail than I can do justice here.

[31] Sterelny and Fraser (2016) offer a defence of such a weaker form of moral realism. I will note that I do not find such approaches plausible, as they commonly rest on a re-definition of what is traditionally understood as moral realism.

[32] This, however, is a matter for another paper. Nagel (2012) recognizes this threat but turns the modus ponens into a modus tollens even going so far as to argue that since moral realism is true the Darwinian story of how morality evolved must be false. This gets things backwards. See Garner (2007) for radical conclusions regarding the elimination of morality, or for nihilism more generally, see Sommers and Rosenberg (2003) and Veit (2019b). A collected volume on the question whether morality should be abolished has recently been published by Garner and Joyce (2019).

science seem to hold a deflated sense of moral objectivity or become error theorists, such as Mackie.

Much more empirical work needs to be done, but the long path to explaining morality is at least partly illuminated by the work of modelers such as Skyrms, Alexander, and others. Clearly, this can only be the beginning of an explanation, but the first steps have been taken. Replicator dynamics have limits and often need to be supplanted with other models, e.g. non-**EGT** models for inheritance and cognitive mechanisms, to provide satisfying explanations of real-world phenomena. A diverse set of multiple models among which as Hegselmann (2009) argues bridges can be built may be the best thing we can hope for, but these as I have argued are importantly not without considerable explanatory power. See also Veit (forthcoming) for a thorough defense of using multiple models, a position I dub 'Model Pluralism'. To conclude, it is just the faulty ideal of a complete explanation that blocks such incremental steps towards a better understanding of complex phenomena such as human morality.

Acknowledgements. First of all, I would like to thank Richard Joyce, Rainer Hegselmann, Shaun Stanley, Vladimir Vilimaitis, Gareth Pearce and Geoff Keeling for on- or offline conversations on the topic. Furthermore, I would like to thank Cailin O'Connor, Caterina Marchionni, and Aydin Mohseni for their comments on a much earlier draft that was concerned with evolutionary game theory models more generally and Topaz Halperin, Shaun Stanley and two anonymous referees for comments on the final manuscript of this paper.

Also, I would like to thank audiences at the 11th MuST Conference in Philosophy of Science at the University of Turin, 2018's Model-Based Reasoning Conference at the University of Seville, the 3rd think! Conference at the University of Bayreuth, the 4th FINO Graduate Conference in Vercelli, the Third International Conference of the German Society for Philosophy of Science at the University of Cologne and the 26th Conference of the European Society for Philosophy and Psychology at the University of Rijeka. Sincere apologies to anyone I forgot to mention.

References

Alexander JM (2007) The structural evolution of morality. Cambridge University Press, Cambridge

Arnold E (2008) Explaining Altruism: A Simulation-Based Approach and its Limits, Ontos

Aydinonat NE (2018) The diversity of models as a means to better explanations in economics. J Econ Methodol 25(3):237–251

Aydinonat NE (2008) The invisible hand in economics: How economists explain unintended social consequences. IN EM advances in economic methodology. Routledge, London

Aydinonat NE (2007) Models, conjectures and exploration: an analysis of Schelling's checkerboard model of residential segregation. J Econ Methodol 14:429–454

Axelrod R (1984) The evolution of cooperation. Basic Books, New York

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396

Binmore K (1994) Playing fair: game theory and the social contract I. MIT Press, Cambridge

Binmore K (1998) Just playing: game theory and the social contract II. MIT Press, Cambridge

Binmore K (2005) Natural justice. Oxford University Press, Oxford

Binmore K, Swierzbinski J, Hsu S, Proulx C (1993) Focal points and bargaining. Int J Game Theory 22:381–409

Boehm C (2012) Moral origins: the evolution of Virtue, Altruism, and shame. Basic Books, New York
Bowles S, Gintis H (2011) A cooperative species: Human reciprocity and its evolution. Princeton University Press, Princeton
Cartwright N (2009) If no capacities then no credible worlds: but can models reveal capacities? Erkenntnis 70:45–58
Churchland PS (2019) Conscience: the origins of moral intuition. W. W. Norton Press, New York
Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
Dennett DC (1996a) Darwin's dangerous idea: evolution and the meanings of life. Simon & Schuster, New York
D'Arms J, Batterman R, Górny K (1998) Game theoretic explanations and the evolution of justice. Philos Sci 65:76–102
D'Arms J (1996) Sex, fairness, and the theory of games. J Philos 93(12):615–627
D'Arms J (2000) When evolutionary game theory explains morality, what does it explain? J Conscious Stud 7(1–2):296–299
de Waal F (1996) Good natured: the origins of right and wrong in humans and other animals. Harvard University Press, Cambridge
de Waal F (2006) Primates and philosophers. Princeton University Press, Princeton
FitzPatrick W (Spring 2016) Morality and evolutionary biology. The Stanford Encyclopedia of Philosophy, Zalta EN (ed). https://plato.stanford.edu/archives/spr2016/entries/morality-biology/
Garner R, Joyce R (2019) The end of morality: taking moral abolitionism seriously. Routledge, New York
Garner R (2007) Abolishing morality. Ethical Theory Moral Pract 10(5):499–513
Gelfert A (2016) How to do science with models. Springer, Dordrecht
Giere R (1999) Science without laws. University of Chicago Press, Chicago
Giere R (1988) Explaining science. University of Chicago Press, Chicago
Godfrey-Smith P (2006) The strategy of model-based science. Biol Philos 21:725–740
Grüne-Yanoff T, Marchionni C (2018) Modeling model selection in model pluralism. J Econ Methodol 25(3):265–275
Guala F (2005) The methodology of experimental economics. Cambridge University Press, New York
Guala F (2002) Models, simulations, and experiments. In: Magnani L, Nersessian NJ (eds) Model-based reasoning. Springer, Boston
Dennett DC (1996b) Darwin's dangerous idea: evolution and the meanings of life. Simon & Schuster, New York
Hausman D (1992) The inexact and separate science of economics. Cambridge University Press, Cambridge
Hegselmann R, Schelling TC, Sakoda JM (2017) The intellectual, technical, and social history of a model. J Artif Soc Soc Simul 20(3):15
Hegselmann R, Will O (2013) From small groups to large societies: how to construct a simulator? Biol Theory 2013(8):185–194
Hegselmann R, Will O (2010) Modelling Hume's moral and political theory - the design of HUME1.0. In: Baurmann M, Brennan G, Goodin R, Southwood N (eds) Norms and values: the role of social norms as instruments of value realisation. Baden-Baden, Nomos, pp 205–232
Hegselmann R (2009) Moral dynamics. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, New York, pp 5677–5692
Hobbes T (1651/1982) Leviathan. Harmondsworth. Penguin, Middlesex

Hume D (2007) A treatise of human nature (ed by Norton DF, Norton MJ). Oxford University Press, Oxford
Hume D (1998) An enquiry concerning the principles of morals (ed by Beauchamp TL). Oxford University Press, Oxford
Joyce R (2006) The evolution of morality. MIT Press, Cambridge
Joyce R (2001) The myth of morality. Cambridge University Press, Cambridge
Kagel JH, Roth AE (1998) Handbook of experimental economics. Princeton University Press, Princeton
Kitcher P (1999) Games social animals play: commentary on Brian Skyrms' Evolution of the Social Contract. Philos Phenomenol Res 59(1):221–228
Knuuttila T (2011) Modelling and representing: an artefactual approach to model-based representation. Stud Hist Philos Sci 42:262–271
Kuorikoski J, Lehtinen A (2009) Incredible worlds, credible results. Erkenntnis 70:119–131
Levy A (2018) Evolutionary models and the normative significance of stability. Biol Philos 33:33
Levy A (2011) Game theory, indirect modeling, and the origin of morality. J Philos 108(4):171–187
Lisciandra C (2017) Robustness analysis and tractability in modeling. Eur J Philsoophy Sci 7:79–95
Mackie JL (1977) Ethics – inventing right and wrong. Penguin Books, Harmondsworth
Mäki U (2005) Models are experiments, experiments are models. J Econ Methodol 12(2):303–315
Muldoon R (2007) Robust simulations. Philos Sci 74:873–883
Nagel T (2012) Mind and cosmos: why the materialist neo-darwinian conception of nature is almost certainly false. Oxford University Press, Oxford
Northcott R, Alexandrova A (2015) Prisoner's Dilemma doesn't explain much. In: Peterson Martin (ed) The Prisoner's Dilemma. Cambridge University Press, Cambridge, pp 64–84
Nowak MA, Highfield R (2011) Super cooperators: Altruism, evolution, and why we need each other to succeed. Free Press, New York
O'Connor C (forthcoming) Methods Models and the Evolution of Moral Psychology. Oxford Handbook of Moral Psychology
Parke EC (2014) Experiments, simulations, and epistemic privilege. Philos Sci 81(4):516–536
Plato (1961) "Protagoras". In: Hamilton E, Huntington C (eds) The collected dialogues of Plato. Princeton UP, Princeton
Rosenberg A (2011) The atheist's guide to reality. W. W. Norton & Company, New York
Rosenberg A, Linquist S (2005) On the original contract: evolutionary game theory and human evolution. Analyze und Kritik 27:137–157
Sadrieh A, Güth W, Hammerstein P, et al (2001) Group report: is there evidence for an adaptive toolbox? In: Gigerenzer G, Selten R (eds) Bounded rationality: the adaptive toolbox, chap 6. The MIT Press, Cambridge, pp 83–102
Seinen I, Schram A (2006) Social status and group norms: indirect reciprocity in a helping experiment. Eur Econ Rev 50(3):581–602
Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge
Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Natl Acad Sci United States 97(16):9340–9346
Skyrms B (1996) Evolution of the Social Contract. Cambridge University Press, Cambridge
Sterelny K, Fraser B (2016) Evolution and moral realism. Br J Philos Sci 68(4):981–1006
Sober E, Wilson DS (1998) Unto others: the evolution and psychology of unselfish behavior. Harvard University Press, Cambridge
Sommers T, Rosenberg A (2003) Darwin's nihilistic idea: evolution and the meaninglessness of life. Biol Philos 18:653–668
Sugden R (2009) Credible worlds, capacities and mechanisms. Erkenntnis 70:3–27
Sugden R (2001) The evolutionary turn in game theory. J Econ Methodol 8(1):113–130

Sugden R (2000) Credible worlds: the status of theoretical models in economics. J Econ Methodol 7:1–31
Street S (2006) A darwinian dilemma for realist theories of value. Philos Stud 127:109–166
Veit W (2019a) Evolution of multicellularity: cheating done right. Biol Philos 34: 34. Online First. https://doi.org/10.1007/s10539-019-9688-9
Veit W (2019b) Existential Nihilism: the only really serious philosophical problem. J Camus Stud 2018:211–232
Veit W (forthcoming) Model Pluralism. Philos Soc Sci. http://philsci-archive.pitt.edu/16451/1/Model%20Pluralism.pdf
Veit W, Dewhurst J, Dolega K, Jones M, Stanley S, Frankish K, Dennett DC (forthcoming) The rationale of rationalization. Behav Brain Sci. PsyArXiv https://doi.org/10.31234/osf.io/b5xkt
Wedekind C, Milinski M (2000) Cooperation through image scoring in humans. Science 288:850–852
Weisberg M (2013) Simulation and similarity: using models to understand the world. Oxford University Press, Oxford
Weisberg M (2007a) Three kinds of idealization. J Philos 104(12):639–659
Weisberg M (2007b) Who is a modeler? Brit J Philos Sci 58:207–233
Williams B (1981) Moral luck. Cambridge University Press, Cambridge
Wilson EO (1975) Sociobiology: the new synthesis. Harvard University Press, Cambridge
Wimsatt W (2007) Re-engineering philosophy for limited beings: piecewise approximations. Harvard University Press, Harvard
Woodward J (2006) Some varieties of robustness. J Econ Methodol 13:219–240
Yaari ME, Bar-Hillel M (1984) On dividing justly. Soc Choice Welf. 1:1–24
Ylikoski P, Aydinonat NE (2014) Understanding with theoretical models. J Econ Methodol 21 (1):19–36
Zollman K (2009) Review of Eckhart Arnold, Explaining Altruism: A Simulation-Based Approach and its Limits. Notre Dame Philosophical Reviews

Coherence and Credibility in the Story-Model of Jurors' Decision-Making: Does Mental Simulation Really Drive the Evaluation of the Evidence?

Marion Vorms[1(✉)] and David Lagnado[2]

[1] University Paris 1 Panthéon-Sorbonne/IHPST (CNRS), Paris, France
marion.vorms@univ-parisl.fr
[2] University College London, London, UK

Abstract. According to the "story-model" of jurors' decision-making, as advocated by Pennington and Hastie (1986, 1988, 1992, 1993), jurors in criminal trials make sense of the evidence through the construction of a mental representation of the events, rather than through the estimation and combination of probabilities. This 'story' consists in a causal explanatory scenario of the crime, and is supposed to drive the jurors' choice of a verdict. As suggested by Heller (2006), the story-model can be described as a legal application of the simulation heuristic (Kahneman and Tversky 1982), according to which people determine the likelihood of an event based on how easy it is to picture the event mentally: the easier to mentally simulate the prosecution scenario, the higher the conviction rate. The primary goal of this paper is to present the main tenets of Pennington and Hastie's (1986, 1988, 1992, 1993) "story-model" of jurors' decision-making, and to draw a few criticisms thereof, in the light of an analysis of evidential reasoning. While acknowledging that some fundamental reasons for adopting this model are well-grounded, and make it a plausible account of jurors' reasoning, we raise some issues concerning its core theses. In particular, we show that the claim that the evaluation of the credibility of the evidence is mediated by story construction, and determined by the coherence of the story, is not tenable as such, and needs to be complemented by a more probabilistically centered approach.

Keywords: Evidential reasoning · Jurors' decision-making · Mental simulation

1 Introduction

Jurors in Common Law criminal trials are requested to bring a verdict based on the analysis of the evidence presented in court. This requires first forming a belief about what happened—jurors are 'fact-finders'—, and then bring a verdict based on the conclusion of that first step (both on the facts they have found, and on whether those

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 103–119, 2019.
https://doi.org/10.1007/978-3-030-32722-4_7

have been established 'beyond a reasonable doubt'[1]). How do they achieve such a complex task? And, how should they? In many regards, the jurors' task can be taken as paradigmatic of evidential reasoning and decision-making under uncertainty, especially where the stakes of the decision are high: jurors face a heterogeneous, incomplete, and partially contradictory set of evidence, upon which they must make highly consequential decisions, without reaching complete certainty about what actually happened.

Evidential reasoning, conceived of as a cognitive activity aimed at drawing inferences from a set of evidence to reach a conclusion about an uncertain issue, can be analyzed along various dimensions. Confirmation theories in the philosophy of science (Hajek and Joyce 2008; Hartmann and Sprenger 2010; Crupi 2016) tend to focus on the relation between a piece of evidence or a datum, and the hypothesis on which it has confirmatory bearing. However, considering the role of evidence in the criminal context draws attention on other, crucial aspects of evidential reasoning (which also have relevance outside the judicial context), and particularly the issue of the credibility of evidence, and the strikingly complex hierarchical relationships among a network of evidential items which may, or may not, cohere (Schum 1994). How do agents navigate such a tangled web of information? And, how should they?

In the psychology of reasoning, there exist (at least) two types of approaches to jurors' decision-making: scenario-based (in particular Pennington and Hastie's 1986, 1988, 1992, 1993), and probability-based (drawing on Schum's 1994 study of evidential reasoning) approaches.[2] Those differ from both a normative, and a descriptive perspective. Whereas probability-based approaches argue that jurors (should) reach a decision by estimating and combining probabilities, scenario-based approaches claim that jurors need to envision a causal scenario of what happened to make their decision. As such, scenario-based views can be considered as a legal application of the simulation heuristics (Kahneman and Tversky 1982), according to which people determine the likelihood of an event based on how easy it is to picture the event mentally (see Heller 2006): the easier to mentally simulate the prosecution scenario, the higher the conviction rate.

One important aspect of the juror's situation is that her reasoning is backward-looking: most of the time, it concerns a particular event (a crime), the exact causal story of which has to be clarified. This is one reason why jurors' reasoning seems particularly well suited to an explanation-based approach. And indeed, among the scenario-based approaches, one of the first, and still most influential proposals, is Pennington and Hastie's (1986, 1988, 1992, 1993) so-called *story-model* of jurors' decision-making, which takes the construction of a coherent, causal explanation, as central. According to it, jurors in criminal trials make sense of the evidence through the construction of a

[1] As a matter of fact, since the turn of the 21st century, the Crown Court has given up the phrase "beyond a reasonable doubt" in its guidance to trial judges in directing the jury, in favor of a mention that "the prosecution must make the jury *sure* that [the defendant] is guilty" to prove guilt, and that "nothing less will do". (Judicial College, *The Crown Court Compendium – Part I: Jury and Trial Management and Summing Up* (February 2017), 5–8 (on line at www.judiciary.gov.uk/publications/crown-court-bench-book-directing-the-jury-2/).

[2] One should also mention argumentative theories of jurors' reasoning; see *e.g.* (Bex 2011) for a "hybrid theory" integrating arguments and stories.

mental representation of the events, rather than through the combination of probabilities. This 'story' consists in a causal explanatory scenario of the crime, and it is supposed to drive the jurors' choice of a verdict.

The story-model is now among the most widely accepted views of jurors' decision-making. It has a strong intuitive appeal: whether, and how easily, one is able to understand what happened by mentally representing the causal sequence of events that led to the death of the victim, seems to be an important component of the very notion of 'reasonable doubt'. In other words, it seems intuitively right that, even in the presence of some strong piece of incriminating evidence, one would feel reluctant to convict someone without being able to construct a coherent and plausible scenario of how then defendant could have committed the alleged crime. Moreover, the model seems to capture important aspects of judicial practice—one of the tasks of the prosecution consists in "story-telling", namely proposing a version of the facts, rather than only exhibiting evidence. This suggests that there should be some normative justification of this, which could be thought of along the lines of theories of inference to the best explanation (Lipton 1991).

This paper is intended to propose a critical clarification of the main tenets of the story-model. Close scrutiny of the story-model reveals important conceptual, as well as empirical, issues, which we shall highlight. In particular, core notions, such as 'coherence', as well as the overall mechanism of story construction, still needs to be clarified. The main claim this paper advocates is that such an approach needs to be complemented by a theory of evidential reasoning that seriously takes into account how evidence is, and should be, analyzed, evaluated, and integrated within one's mental model. In other words, we aim at overcoming the opposition between coherence, and probability-based approaches, which is mostly grounded on caricatural pictures of each. Although probability-based approaches are mainly normative, and the most reasonable reading of the story-model is descriptive, we aim to show that the story-model suffers conceptual and empirical problems that jeopardize it as a descriptive account, and that complementing it with an account of evidential reasoning which is compatible with (normative) probabilistic models makes it more plausible as a descriptive account—and opens up the perspective of a tractable normative program. We thereby hope to contribute to a more general reflection on the articulation between considerations of coherence, which are characteristics of scenario, and explanation-based approaches, and attention to the credibility, relevance, and probative strength of evidence. This paper will mostly be critical in its argumentation, but should be taken as the 'negative' part of a constructive project involving an experimental program, which we sketch at the end of the paper.

2 Pennington and Hastie's 'Story-Model': Core Theses and Experimental Evidence

The following presentation of the story-model is based on a series of theoretical and experimental papers by Nancy Pennington and Reid Hastie (1986, 1988, 1992, 1993). We first present what we take as the two core theses of the model, conceived of as a psychological theory of jurors' decision-making (Sect. 2.1). We then briefly report

some key experimental evidence backing those two claims (Sect. 2.2), after which, we introduce some additional notions intended at giving some flesh to the main two claims —but whose empirical meaning is less clearly explicated (Sect. 2.3). Some experimental results aiming at clarifying such meaning are then reported (Sect. 2.4); those results are taken as a support for a third, most controversial, thesis, which we finally expose (Sect. 2.5).

2.1 Core Theses

Pennington and Hastie's story model essentially consists of two fundamental claims. Those respectively correspond to the main two steps of the juror's task, namely *i.* making sense of the evidence so as to form a belief regarding what happened, and *ii.* bringing a verdict supposedly based on that belief, and meeting the appropriate standard of proof.

Thesis #1. Evidence Processing as Story Construction. The main, fundamental claim by Pennington and Hastie's is that jurors process trial evidence by constructing a causal, mental model of what happened—they mentally simulate the chain of events that led to, say, the death of the victim. Such construction is not additional to, or a consequence of, evidence interpretation; rather, it is how jurors make sense of the evidence. "During the course of the trial, jurors are engaged in an active, constructive comprehension process in which they make sense of trial information by attempting to organize it into a coherent mental representation (1992, 190)." This results in a "mental representation of the evidence that constitutes an interpretation of what the evidence is about" (1988, 521).

Such mental representations, called 'stories', involve chains of physical, as well as psychological events (crime stories typically involve goals, intents and motives). Stories themselves are composed of sub-stories called 'episodes', which are made of causal links between physical and mental events. It is important to insist that stories, as mental representations, are made of *events*, rather than of the *evidence* that is supposed to attest to such events.

Story construction can typically be prompted or facilitated by the prosecution address, which most often consists in organizing evidence so as to make it fit a particular causal sequence of events that led to the alleged crime. But Pennington and Hastie's claim goes further: they claim that jurors *spontaneously* impose a story-structure on the evidence. Evidence processing is not dissociable from mentally simulating what happened.

Since evidence presented at trials is seldom exhaustive—it hardly provides all the elements that are needed to tell a complete story, Pennington and Hastie insist that stories "incorporat[e] inferred events and causal connections between events in addition to relevant evidentiary events." (1988, 521). And in fact, this will be a crucial aspect of their interpretation of experimental data as supporting their model, as exposed below.

Thesis #2. Verdict Categories and Verdict Choice. The second, main claim by Pennington and Hastie's is that story construction determines verdict choice, by allowing jurors to process the evidence "such that evidence can be meaningfully evaluated against multiple verdict judgment dimensions" (1992, 192). This is done

through a matching of the structure of the story with the criteria of the verdict, as enunciated by the judge. Quite notably indeed, legal verdict categories correspond to human action sequences, like stories: they explicitly refer to mental states such as intentions and goals. For example, for the qualification of "first degree murder", there must be intent to kill, whereas there is no intent in a manslaughter.

In summary, during trials, jurors interpret evidence by constructing a mental representation of the events, and then compare the causal structure of that representation with the typical structure of verdicts, so as to come to a decision.

2.2 Empirical Evidence Backing Claims #1 and 2—Interview Studies

Pennington and Hastie's (1986) original experimental procedure consisted in interview studies based on a verbal protocol. Having been presented with some trial materials, mock jurors were asked to talk aloud while making their decision, and to respond to questions about the evidence and about the instructions they got from the judges. The goal of such experiment was to "elicit data that would provide a snapshot of the juror's mental representations of evidence and of verdict categories at one point in time" (1993, 203).

The analysis of the verbal data so obtained revealed that 85% of the events described by the subjects were causally linked (*e.g.* "Johnson was angry so he decided to kill him"), which Pennington and Hastie took as "strong evidence that subjects were telling stories and not constructing arguments" (1993, 205). They thus concluded that mental representations of the evidence "showed story structure and not other plausible structures" (1993, 205).

Another important result from those original studies was that "only 55% of the protocol references were to events that were actually included in testimony. The remaining 45% were references to inferred events—actions, mental states, and goals that 'filled in' the stories in episode configurations." (1993, 206). As mentioned above, one aspect of the story-model is that stories need to be richer representations than what is warranted by the evidence. According to Pennington and Hastie, this result "argues strongly against the image of the juror as a "tape recorder" with a list of trial evidence in memory" (1993, 206).

In support of the second main thesis presented above, experimental data reveal that people construct different stories from the same evidence, and that "story structures differed systematically for jurors choosing different verdicts." (1993, 206) Lastly, such link between story structure and verdict is interpreted as a causal relation between the former and the latter. To rule out the hypothesis that story construction may be a post hoc rationalization for the choice of a verdict (rather than a determinant thereof), Pennington and Hastie (1988) gathered evidence that "mental representations that resemble stories are constructed *spontaneously* in the course of the juror's performance, that is, without the prompt of an experimenter's requests to discuss the evidence or the decision" (1988, 523).

2.3 Story Acceptance and Verdict Choice

Two steps are described in the model just presented. The first one consists in the construction of a story. The second one is the choice of a verdict. However, several stories can be constructed on the basis of the same evidence; but, at the end of the day, the juror must select one of them, as it is supposed to provide the ultimate ground for verdict choice. Let us see in more details what Pennington and Hastie tell us about the criteria governing those two decisional steps.

Story Acceptance: Certainty Principles and Levels of Confidence. Evidence does not uniquely determine a story. And in fact, one common strategy for defense lawyers is to propose an alternative, exonerating scenario—although they do not have to do so, and are allowed just to raise issues about the prosecution's scenario without constructing an alternative one[3]. Besides, different jurors may come up with different stories, and the very same juror can herself imagine different scenarios. One story, however, has to 'win'. Who the 'winner' will be depends on how well it satisfies a set of criteria, called 'certainty principles' (1992, 190). Those principles both determine which story is the most acceptable, and how much confidence the juror will place in it, once it is accepted[4]. As we will see, such relative confidence also has a bearing on verdict choice.

Let us now see what certainty principles consist in.

Coverage and Coherence. The first two principles enunciated by Pennington and Hastie are *coverage* and *coherence*. 'Coverage' corresponds to how much the story covers the available evidence. 'Coherence', on the other hand, is defined in terms of three components: *a.* consistency (how little contradiction there is among the elements of the story), *b.* plausibility (how much it matches background knowledge and common assumptions about how the world goes, how people act in general, etc.), *c.* completeness ("the extent to which the story has all of its parts", 1992, 191). Coverage and coherence together determine both the acceptability of a story, and the confidence one places in it, once accepted.

Uniqueness. To these first two criteria, Pennington and Hastie add a third one, namely *uniqueness*: one unique, good, story is better than two rival ones. This principle, which does not bear on the intrinsic characteristics of a story but rather on the context in which it is constructed, cannot be used as a measure of its acceptability. However,

[3] There are different views as to what might be the best strategy. McKenzie *et al.* (2002) research (as reported by Heller 2006, 263) into the relationship between prosecution and defense cases suggests that a defense case reduces confidence in the prosecution's case only if it exceeds its "minimum acceptable strength", a threshold that is determined by the strength of the prosecution's case. And he finds that a defense case can backfire if it fails to exceed its minimum acceptable strength. Heller reinterprets McKenzie's results in the framework of the story-model, through the notion of 'ease of imagination'.

[4] Pennington and Hastie are not very specific about their notion of acceptance, and how it relates with levels of confidence. It seems that they consider acceptance as an all-or-nothing matter—that at some point one accepts one story and rejects the others, based on a comparison of their relative satisfaction of the various certainty principles—, but that, once accepted, it can be held with different levels of confidence.

when there are more than one theory that is judged acceptable, "great uncertainty will result" (1992, 191), which should impact verdict choice. The interpretation of this principle is not entirely clear, and this raises several philosophical, as well as empirical issues that deserve close attention, but are beyond the scope of this paper.[5]

Verdict Choice: Goodness of Fit, and the Standard of Proof. How does the acceptance of a given story determine the choice of a verdict? As mentioned above, verdict categories (first degree murder, manslaughter, etc.) typically correspond to human causal actions sequences, like stories. The second step of the juror's task is therefore to assess whether, and how much, the structure of the story she has accepted matches the structure of one of the verdict categories presented by the judge (see Fig. 1).

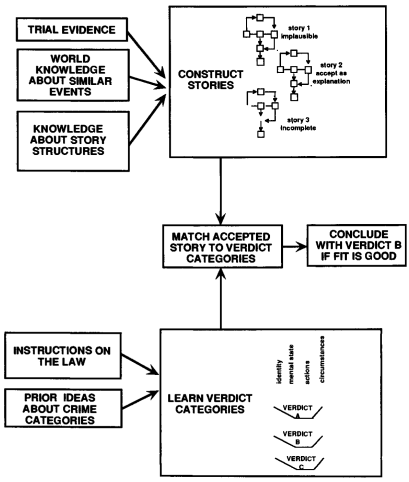


Fig. 1. The story-model for juror decision-making (Pennington and Hastie 1993, 193)

Goodness of Fit. One first thing is to determine the "best-match verdict" (1992, 192), namely the verdict category whose structure is the closest to the one of the accepted story. But then one also needs to assess how nicely the two match together. This is measured through the fourth 'certainty principle' introduced by Pennington and Hastie (1992, 1993), namely 'goodness of fit'. This latter principle is supposed to govern—together with the other three—"the confidence or level of certainty with which a particular decision will be made" (1993, 193). How it does so is not entirely clear, though, as we will now briefly see.

The Question of the Standard of Proof. As is well known, jurors in criminal trials are instructed to bring a verdict of guilty when guilt has been proven 'beyond a reasonable

[5] This is the object of our experimental work in progress with Katya Tentori and Saoirse Connor Desai.

doubt'.[6] There are important debates as to what the right interpretation of the standard is, but no consensual definition has been reached in legal theory or practice.[7] By contrast with interpretations of the standard in terms of a probabilistic threshold, Pennington and Hastie describe it in terms of goodness of fit: "If the best fit is above a threshold requirement, then the verdict category that matches the story is selected. If not all of the verdict attributes for a given verdict category are satisfied 'beyond a reasonable doubt' by the events in the accepted story, then the juror should presume innocence and return a default verdict of not guilty. We are basing this hypothesis on the assumption that jurors either (A) construct a single 'best' story, rejecting other directions as they go along or (B) construct multiple stories and pick the 'best.' In either case, we allow for the possibility that the best story is not good enough or does not have a good fit to any verdict option. Then a default verdict has to be available" (1993, 201).

Pennington and Hastie here seem to suggest that both the level of confidence one places in the story, and the goodness of fit of that story with the verdict category matter to determine whether the corresponding charge has been established beyond a reasonable doubt. Although this may seem intuitively right, several issues arise about this view. Firstly, no precise indication is given about how the different measures (confidence, match between events and 'verdict attributes') can be operationalised, nor about how they trade off with each other. Second, and more importantly, it is not clear whether this view is supposed to be normative or descriptive. If descriptive, what empirical evidence could there be in its support? It is hard to see how the several aspects of this complex construct might be tested. Similarly, if the view is normative, then we need more precise measures of fit to turn this into an operational definition of reasonable doubt.[8] As such, the model provides no clear criteria for either the required level of confidence in the story or for the goodness of fit with a verdict category. We will leave these issues aside, as they are beyond the scope of the present criticism.

To those issues, however, additional problems arise with regard to the empirical meaning and testability of what can be taken as the most important certainty principle, namely coherence. Let us now focus on this.

2.4 Experimental Manipulation of 'Ease of Construction': Presentation Order

No clear and operational definition is given of the various certainty principles, and nothing clear is said about how they interact in the juror's belief formation and decision-making processes. In particular, 'coherence' is far from having a clear and consensual definition, although it is central to the model. However, experiments by Pennington and Hastie (1988, 1992) suggest a way to indirectly test the impact of story coherence on confidence and verdict choice, through a manipulation of the 'ease of

[6] But see footnote 1.

[7] See *e.g.* Laudan 2008, 32–37, Roberts and Zuckermann 2010, 253–258.

[8] See Laudan 2007 for a criticism of Inference of the Best Explanation as a candidate to understand the BARD ('Beyond a Reasonable Doubt') standard.

construction' of stories. The basic idea is that, the easier to construct, the more complete and therefore coherent the story will be, which should result in a higher confidence in the story, and hence in the verdict choice.

Pennington and Hastie's experimental protocol consist in influencing story construction (both which story will 'win', and how strongly it will be accepted), by varying the presentation order of the evidence—but not its content[9]—, and in testing the effects of such manipulation on the verdict choice and confidence. It is important to note that the effect here is not supposed to be an order effect (according to which depending on whether it was presented first or last a piece of evidence will not have the same weight). One important aspect of Pennington and Hastie's model is that evidence evaluation is a holistic, rather than sequential, process.

There are obviously several non-trivial assumptions here, which we will not scrutinize in this paper. One core assumption though is that stories are easier to construct when evidence is ordered in a temporal and causal sequence that matches the original events (what they call "story order"). What counts as "no story order" is not entirely clear though: in their 1988 paper, it corresponds to witness order (the order in which witness come to the bar to give testimony), but, in 1992, the witness order actually corresponds to the 'story order', and the 'no story order' corresponds to the presentation of the evidence issue by issue. Such difference might be explainable in terms of the content of the reports themselves; but this calls for clarification. And, empirically, what may count as story order is an empirical question that needs further inquiry.

In their 1988 experiments, Pennington and Hastie found that, when evidence is presented in story order, *a.* subjects make more decisions in the direction of the preponderance of the evidence (they are likeliest to convict when presentation order facilitates the construction of a prosecution story, and vice-versa), and the conviction rate is greater: "Thus, story coherence, as determined by presentation order of evidence, affects verdict decisions in a dramatic way" (1993, 211). *b.* They express more confidence in those decisions when the evidence is presented in the story order: "the coherence of the explanation the decision maker is able to construct influences the confidence associated with the decision." (1988, 521)

In another study (1992), they replicate this effect (although 'story order' is implemented differently, see above). Subjects are presented with three consistent witnesses' reports, and a fourth one which is inconsistent with the first three. The three consistent witnesses are presented as credible. The fourth one's credibility varies. Results show that credibility information about a source of testimony has more impact in the 'story order condition'. From this second study, Pennington and Hastie draw a conclusion, which essentially corresponds to what can be considered as their third, core thesis, namely that "ease of story construction mediates perceptions of evidence strength, judgments of confidence, and the impact of information about witness credibility." (1992, 202)

[9] That manipulating presentation order without modifying the content is possible is no trivial assumption, though.

2.5 Core Thesis #3 Story Coherence and Evidential Value

Although we are left in the obscurity about the very meaning of this 'mediation', it is at the core of what can be taken as the third core thesis by Pennington and Hastie, which is by far the most controversial and problematic one, namely that *evidence organization, interpretation, and processing is mediated through story construction*. What does that mean?

Pennington and Hastie insist on the "distinction between the evidence presented at trial and [their] concept of verdict stories. The evidence is a series of assertions by witnesses that certain events occurred. The juror chooses to believe only some of these assertions and will make inferences that other events occurred, events that are never stated explicitly in the evidence, as well as inferences concerning causal connections between events. This activity on the part of the juror results in a mental structuring of the evidence that is the 'juror's story.' (Pennington and Hastie 1988, 524) Hence, although evidence presented at trial provides the main ground for story construction, it does not constrain it in any strong way: jurors may select the evidence items they judge relevant for their story, and complement them with the representation of events inferred from their background knowledge. Such additional inferences are particularly important to fill in psychological events such as goals, intentions, and beliefs.

Based on the experimental results presented above, Pennington and Hastie sketch what could be taken as their own theory of evidential reasoning, by claiming that the very evaluation of the evidence—evaluation of its credibility, relevance, and probative force, which is central to any theory of evidential reasoning—is dependent upon story construction. "The basic claim of the story model is that story construction enables critical interpretive processing and organization of the evidence." (1993, 203)

As explained above, evidence assessment must finally be dependent upon the coherence of the story, and on how well a given piece of evidence fits that story. Coherence, and explanatory power, are the main virtues a story must have to be acceptable. And it is the story itself which determines whether a piece of evidence should be taken into account, or not, as the "perceived importance of a piece of evidence is determined by its place in the explanatory structure imposed by the decision maker during the evidence evaluation stage of the decision process." (1988, 527)

Hence, evidential strength does not so much depend on the evidence presented at trial, than on the virtues of the story itself: the "perceived strength of the evidence for or against a particular verdict decision is a function of the completeness of the story constructed by the juror" (1992, 196). And in turn, it is the structure and characteristics of the story itself, which drive evidence evaluation, and not the other way around: "Story coherence as determined by the presentation order of evidence affects perceptions of evidence strength, verdict decisions, and confidence in decisions" (1988, 529). This is a highly controversial claim, both from a normative, and from a descriptive point of view.

3 A Theory of Evidential Reasoning? Limits of the Model

One of the declared goals of the model is to provide "a psychological account of the assignment of relevance to presented and inferred information" (Pennington and Hastie 1993, 203). It therefore presents itself as a theory of evidential reasoning, explaining how people deal with complex evidence, when having to make a decision based thereupon. But, as emphasized above, it consists in making evidence evaluation conditional on story construction. This claim calls for a thorough examination, both from a normative, and from an empirical point of view.

3.1 Conceptual Issues

No one would deny that verdict choice, insofar as it is supposed to depend on the conclusions of the fact-finding process, should be grounded on the evidence presented at trial. How does the story-model accommodate this rather uncontroversial (normative) assumption? What does it tell us about the link between judicial evidence and verdict choice?

As we have just seen, it is the intrinsic virtues of the story itself (how much it satisfies the acceptability principles), rather than of the evidential set, which provide strength to the confidence placed in it, and hence in the subsequent verdict. In other words, the story, rather than the evidential set on which it is supposed to be constructed, bears evidential force for a given verdict: it is the story that provides reasons to choose it. But, as we have seen, the story is not uniquely determined by the evidential set: quite the opposite, what evidence will actually be taken into account by the juror depends on how well it fits within the story.

Pennington and Hastie would surely not maintain that the decision-making process is disconnected from the trial evidence: after all, story construction is supposed to be prompted by evidence. However, it is far from clear what the functional relationship is, between the story supposedly constructed on the basis of a set of evidential items, and such items whose relevance, credibility, and probative force is assessed through story construction.

Evidence and Events. As mentioned above, Pennington and Hastie take as crucial the "distinction between the evidence presented at trial and [their] concept of verdict stories" (1988, 524). Indeed, their view is that only some pieces of evidence feature in the stories, which are complemented by further inferences. To make our point clear in the following, it might be useful to introduce a distinction proposed by Schum (1994), between a piece of evidence E*—most often a report, either by a lay or by an expert witness, but E* can also refer to any physical evidence presented at trial—, and the fact or event E that E* attests to. Where E* is the victim's neighbor's testimony that she

saw the defendant in the staircase five minutes before the hour of the alleged crime, E is the fact that the defendant was actually in the staircase at that moment[10].

In Schum's terms, what Pennington and Hastie's stories represent is E, not E*. The objects of jurors' mental simulation are the events E_n themselves, without consideration of the evidential items E_n*, which do not feature as such in the model. Of course, they may be part of the model insofar as, taken as events themselves, they are causally connected to the main events. Consider for example a case where experts report that some DNA sample was found on the crime scene that matches the defendant's DNA: this reported fact might feature as an effect of the defendant's actions as represented in the story. Similarly, the fact that the witness saw the defendant in the staircase might feature as an event related to the actual presence of the defendant. But the experts' and witness' reports in court, as evidence for those facts, are not part of the crime story. To be sure, a juror's mental representation may be rich and encompassing enough to feature aspects of the trial as part of the crime story: after all, the experts' as well as the witness' testimonies are themselves events that are causally related to the events featuring the crime. However, where a juror's story includes such events, it does not involve any reasoning process on the evidence as such—on its credibility, relevance, taken independently from the story in which it may, or may not, feature.

In brief, presented with a series of evidential items E_1*, E_2*, … E_n*, jurors either represent E_1, E_2, … E_n, or not, but nothing is said in the model about their entertaining E_1*, E_2*, … E_n* as such—about considering whether, and how much, they are credible and relevant. If they feature in the mental simulation, it is not as evidence items, but rather as events themselves. Of course, their explanatory power is part of the construction of a good causal model—and in fact, as we will mention later, it is likely that jurors do construct such complex causal models including evidence items. But, as such, the story-model leaves us in the dark as to how the evidence items are selected in the first place—or how the story is constructed in the first place.

As mentioned earlier, Pennington and Hastie are quite aware of this distinction, since they insist that their theory is that jurors' reasoning primarily concerns the events —what the evidence is about: "On the basis of this research, we assume that when evidence is presented, the subject constructs a verbatim representation of the surface structure of the evidence (in the present experiment the evidence was presented as a written text), a semantic representation of the evidence in the form of a propositional textbase, and a situation model that represents an interpretation of what the evidence is about (in our terms, this is the juror's story that we hypothesize is a determinant of the juror's decision)." (1988, 524).

So stated, the story-model appears as very different from accounts on evidential reasoning that take as a central task evaluation of credibility and probative force of evidence (such as the Bayesian approach, and more generally any probabilistic approach). Indeed, as we have just seen, according to the story-model, what counts as a

[10] To be more precise, we should add in an intermediary fact, namely she saw the defendant in the staircase, from which we might infer that he was actually there. As Schum argues, such chains of inferences can be indefinitely decomposed. See his chapter 3 about the components of the credibility of a testimonial report.

reason to accept a story, and choose a verdict, is only indirectly related to credibility and strength of evidence. Rather, it is the story which is the bearer of the evidential strength, in support of a verdict.

As such, this might not be taken as a strong objection to the model: quite the contrary, one of the findings put forward by the advocates of the story-model is that, contrary to what probabilistic approaches claim, the evaluation of evidence credibility and strength is *mediated* through story construction. However, the lack of clarity on how the whole process of story construction works, and on what such models consist in, raises further issues.

Circularity. As a symptom of such issues, let us start by mentioning that Pennington and Hastie are sometimes unclear about whether the story is an explanation of the facts, or of the evidence: "decision makers begin the decision process by constructing a causal model to explain the available facts" (1988, 521), but "the decision maker constructs a causal explanation of the evidence" (1988, 521).

This might well be due to the fact that stories have a double mediating function: governing evidence evaluation, and being a mediation between such evidence and the choice of a verdict. Why, and how, may jurors "choose to believe some assertions rather than others"? There seems to be some circularity in the process as described by Pennington and Hastie: where does the explanatory structure come from in the first place, if not from the consideration of the evidence, and the evaluation of its importance? What drives the choice of a given piece of evidence? How is evidence chosen for the construction of a story, if it is the story which drives evidence evaluation? So, if story construction drives evidence processing, what drives story construction, and the choice of a particular one among the various that may be constructed to make sense of the evidence? In other words, if jurors' mental representation is on E, how do they initially infer E from E* (if the credibility of different E_s* depends on which E_s feature in their story)?

The 'certainty principles' are not well defined enough to figure out precise answer to those questions. True, the 'coverage' principle concerns the relation between evidence and story, but it lacks any operational definition, and it is not clear how it may trade off with other principles. In brief, the functional relationship between evidence and events is unclear, and that threatens the explanatory, and predictive power of the model.

Credibility and Coherence Are Not Independent. Let us now come to what we see as the major limitation of the model. As stated above, the central claim by Pennington and Hastie, which can be taken as the core principle of their theory of evidential reasoning, is that the credibility of a piece of evidence is evaluated through story construction, and that the more coherent the story, the more sensitive jurors are to the credibility of a piece of testimony.

It is worth insisting that coherence is coherence of the story itself, as mentally elaborated and represented (how well it fits together, and with the jurors' background knowledge). It is not coherence of the evidential set (how little contradiction there might be between the evidentiary items presented at trial). However, the way the evidential set is presented, as we have seen, may impact the coherence of the story through facilitating—or preventing—story construction. The fact that Pennington and

Hastie's own definition of coherence is not entirely explicit might not be such a big problem, since it can be supplemented by other approaches to explanatory coherence, such as Thagard (2000), Simon and Holyoak (2002). And in fact, Byrne (1994) has attempted to show the convergence of Thagard's views on explanatory coherence, and Pennington and Hastie's model.

However, there is some further worry with the claim that the evaluation of the credibility of a piece of evidence (as well as its relevance, and strength) depends on story coherence. On any precise account of model coherence, coherence of the evidential set has a role to play. But the credibility of a piece of evidence cannot be taken as independent from its coherence with other pieces of evidence, which themselves attest to other events. In fact, one important lesson from Schum's structural analysis of evidence, and from the legal scholarship in the line of Wigmore's (1937), which often uses Bayes nets to formalize inference chains, is that whether an agent chooses to believe such or such source, and hence use such or such piece of evidence for the construction of her story, has consequences on the credibility of other pieces of evidence, and hence on the coherence of the whole set (see Lagnado and Harvey 2008; Lagnado 2011). An evidential set is not a collection of items that can be considered separately, but rather a complex, hierarchical network, with subtle and multidimensional internal dynamics. Works by formal epistemologists provide precise, quantitative models of how coherence and credibility may interact in the evaluation of multiple testimonies (Bovens and Hartmann 2003).

This is not to deny that constructing causal models of the crime is the right strategy to deal with such a complex evidential network, but that the causal model cannot ignore the internal dynamics of the evidential set itself. It is not clear what the coherence of a piece of evidence within a story means, independent from other considerations. Consider a set of evidential items. Where there is contradiction between some of them, this may lower the credibility of each of them. How do jurors deal with contradictory items? What makes them choose one, rather than the other?

In other words, evaluating a piece of evidence (say, a testimonial report) as not credible is non dissociable from telling a story about how it was produced (by suggesting that the witness was interested in such or such outcome of the trial, or that she has some memory problems, etc.). A more complete theory of story construction should take account of this. But Pennington and Hastie rather seem to consider that jurors simply dismiss pieces of evidence that do not fit their story, without providing any causal explanation for their existence.

Hence, the claim that "the Story Model directly addresses the question 'Where do the weights come from?'" (1988, 527) does not sound totally legitimate: the story-model does not allow for a clear account of selection of evidence and assessment of its credibility.

3.2 Empirical Adequacy and Completeness of the Model

One could argue that the issues raised above are problematic if one wants to provide a normative model of jurors' decision-making, but that this does not jeopardize the model as a descriptive one. After all, it is possible to argue that jurors spontaneously and rather idiosyncratically generate some representation of a series of events on the

basis of an a-rational consideration of the evidential set, together with the prosecution's address, and then re-consider each piece of evidence in the light of this spontaneous representation. That this is not epistemologically desirable is another issue.

However, at least two kinds of issues arise as to the empirical adequacy of the model. The first one is that it is not even clear what empirical evidence there could be in favor of such an underdetermined model. Indeed, as soon as one tries to design some experimental protocol aiming at testing whether subjects are indeed more sensitive to evidence credibility or to story-order, for instance, one realizes that it is practically impossible to manipulate one without impacting the other.

Second, even empirically, the model needs to be complemented. It is not only normatively true that jurors *should* reason about witnesses' credibility before accepting their report (that it is not only the way they fit in a coherent story—whose origin is not clear—that drives their acceptability). There is empirical evidence that this is what they do: as shown by Connor Desai *et al.* (2016), people do draw inferences about credibility of witnesses, and their motivations. They construct a causal 'story of the trial', in complement to the 'story of the crime', and the two interact in complex ways.

Moreover, there is now robust evidence that agents correctly deal with the dynamics of coherence and credibility, at least qualitatively (Harris and Hahn 2009; Lagnado 2011; Lagnado and Harvey 2008). The story-model still needs to be complemented to accommodate that. Such considerations are not aimed at dismissing the story-model altogether. Rather, we claim that it is perfectly compatible with probability-based approaches, which are too often caricatured as describing human agents as super calculators. Without claiming that jurors should (and could) compute complex Bayesian calculations, the use of qualitative Bayes nets (Lagnado *et al.* 2013) seems to be a promising path to represent how aspects of evidence credibility are taken into account into the computation of causal models.

The story-model thus needs to be complemented by an account of evidential reasoning. Story construction as such cannot provide a complete account. One needs to account for how subjects analyze different items of evidence (witness testimonies, expert reports, etc.), and how this affects story construction and evaluation.

3.3 Accounting for the Variety of Types of Evidence—How Do Jurors Process Forensic Evidence?

As is well-known, judicial evidence can be of various sorts, from expert reports to lay witness testimony, and physical objects, recordings, written documents, etc. It would be rather risky, we suspect, to assume that those diverse types of evidence are similarly analyzed by jurors. It is now common knowledge that eyewitness testimony is rather unreliable, and that scientific evidence is to be taken more seriously—though cautiously as well, for other reasons having to do with the communication of scientific results and the use of statistics. However, how do people, in practice, deal with those different types of evidence? One implication of the story-model is that, the more narratively the evidence is presented at trial, the more confident jurors will be in reaching their verdict. For those reasons, testimonial evidence is likely to have more impact, as it is intrinsically narrative (see Heller 2006).

However, is it really the case that jurors are more influenced by a piece of verbal testimony than by a scientific report? Even though the need for narration is well documented—and rather intuitive, it seems highly dubious that, for instance, a strong exculpatory forensic report should have no weight against an otherwise coherent, but weak, set of incriminating lay testimony.

These are empirical questions, which call for experimental testing. One important project would be to test the relative influence of a piece of forensic evidence (by manipulating its strength) in comparison with a story (manipulating its coherence by varying presentation order following Pennington and Hastie's protocol). Does a coherent story with weak forensic evidence trump a less coherent story with strong evidence? To what extent? How much does the strength of a piece of forensic evidence matter, depending on whether the rest of the evidence is narrative or not? How much does it trump the story when it goes in the other direction?[11]

4 Conclusion

Although it was first proposed in the early 1980s, the story-model is still the most influential account of jurors' reasoning and decision-making. We suspect that this is because of its strong intuitive appeal—which is most probably a sign that it accurately captures something of how jurors make sense of complex evidence. Moreover, we fully agree that that Pennington and Hastie's "explanation-based approach could be viewed as complementary to these other models [information integration and Bayesian models]." (1988, 531) However, we do not see such complementarity as they do: rather than providing "an account of which conditional dependencies between evidence items will be considered in Bayesian calculations", we claim that the story-model's main virtue is to draw our attention to the importance of causality in mental simulation, and that explanation-based reasoning should be supplemented with a framework for evidence evaluation, as provided *e.g.* by Bayes nets (Lagnado *et al*. 2013).

References

Bex FJ (2011) Arguments, Stories and Criminal Evidence, vol 92. Law and Philosophy Library. Springer, Cham

Bovens L, Hartmann S (2003) Bayesian Epistemology. Oxford University Press, Oxford

Byrne M (1994). http://chil.rice.edu/byrne/Pubs/git-cs-94-18.pdf

Connor Desai S, Reimers S, Lagnado DA (2016) Consistency and credibility in legal reasoning: a Bayesian network approach. In: Papafragou A, Grodner D, Mirman D, Trueswell JC (eds) Proceedings of the 38th annual conference of the cognitive science society, Austin, TX, pp 626–631

Crupi V (2016) Confirmation. In: Zalta EN (ed) The stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2016/entries/confirmation/

[11] This is the object of our experimental project with Saoirse Connor Desai and Katya Tentori.

Harris A, Hahn U (2009) Bayesian rationality in evaluating multiple testimonies: incorporating the role of coherence. J Exp Psychol 35(5):1366–1373
Hajek A, Joyce J (2008) Confirmation. In: Psillos S, Curd M (eds) Routledge companion to the philosophy of science, New York, pp 115–129
Hartmann S, Sprenger J (2010) Bayesian epistemology. In: Bernecker S, Pritchard D (eds) Routledge companion to epistemology, London, pp 609–620
Heller KJ (2006) The cognitive psychology of circumstantial evidence. Mich Law Rev 105 (2):241–306
Kahneman D, Tversky A (1982) The simulation heuristic. In: Kahneman D, et al (eds) Judgment under uncertainty: heuristics and biases
Lagnado DA, Harvey N (2008) The impact of discredited evidence. Psychon Bull Rev 15 (6):1166–1173
Lagnado DA (2011) Thinking about evidence. In: Dawid P, Twining W, Vasilaki M (eds) Evidence, inference and enquiry. OUP, British Academy, pp 183–223
Lagnado DA, Fenton N, Neil M (2013) Legal idioms: a framework for evidential reasoning. Argum Comput 4:46–53
Laudan L (2007) Strange bedfellows: inference to the best explanation and the criminal standard of proof. University of Texas Law, Public Research Paper No. 143
Laudan L (2008) Truth, error, and criminal law: an essay in legal epistemology. Cambridge studies in philosophy and law
Lipton P (1991) Inference to the best explanation. Routledge, London
McKenzie CRM, Lee SM, Chen KK (2002) When negative evidence increases confidence: change in belief after hearing two sides of a dispute. J Behav Decis Making 15(1):1–18
Pennington N (1986) Hastie R (1986) Evidence evaluation in complex decision making. J Pers Soc Psychol 51:242–258
Pennington N, Hastie R (1988) Explanation-based decision making: effects of memory structure on judgment. J Exp Psychol Learn Mem Cogn 14:521–533
Pennington N, Hastie R (1992) Explaining the evidence: tests of the story model for juror decision making. J Pers Soc Psychol 62:189–206
Pennington N, Hastie R (1993) The story model for juror decision making. In: Hastie R (ed) Inside the juror: the psychology of juror decision making. Cambridge University Press, Cambridge
Roberts P, Zuckermann A (2010) Criminal evidence, 2nd edn. Oxford University Press, Oxford
Schum D (1994) The evidential foundations of probabilistic reasoning. Wiley, New York
Simon D, Holyoak KJ (2002) Structural dynamics of cognition: from consistency theories to constraint satisfaction. Pers Soc Psychol Rev 6:283–294
Thagard P (2000) Coherence in thought and action. MIT Press, Cambridge
Wigmore JH (1937) The science of judicial proof, as given by logic, psychology, and general experience, and illustrated in judicial trials. Little, Brown



Insight Problem Solving and *Unconscious Analytic* Thought. New Lines of Research

Laura Macchi(✉), Veronica Cucchiarini, Laura Caravona, and Maria Bagassi

Department of Psychology, University of Milano-Bicocca, Milan, Italy
laura.macchi@unimib.it

Abstract. Several studies have been interested in explaining which processes underlie the solution of insight problems. Our contribution analyses and compares the main theories on this topic, focusing on two contrasting perspectives: the *business-as-usual* view (conscious and analytical processes) and the *special process* view (unconscious automatic associations). Both of these approaches have critical aspects that reveal the complexity of the issue on hand. In our view, the insight problem solution derives from an *unconscious analytic thought*, where the unconscious process is not merely associative (by chance), but is achieved by a covert thinking process, which includes a relevant, analytic, goal-oriented search.

1 Introduction

A large body of research has been interested in explaining which processes underlie the solution of insight problems. Their study provides a privileged route to understand creative thought, scientific discovery and innovation, and all situations in which the mind has to face something in a new way. Our contribution aims to present a brief excursus on the main theories in literature on this topic, in particular, focusing on two contrasting perspectives: the *business-as-usual* view and the *special process* view. For the former, the process underlying the resolution of insight problems is conscious and analytical, while for the latter, it is unconscious and associative. Both, studies on the Working Memory Capacity, and on the incubation effect have reported different and often contrasting results. According to our proposal, which will be illustrated in the following pages by referring to the most recent experimental evidence, the insight problem solution derives from an *unconscious analytic thought.* In our view, the insight problem solution is achieved by a productive, creative thinking, resulting from a covert and unconscious thinking process, and an overall spreading activation of knowledge, which includes a relevant, analytic, goal-oriented search.

2 Insight vs. Non-insight Problems

"A problem arises when a living creature has a goal but does not know how this goal is to be reached. Whenever one cannot go from the given situation to the desired situation simply by action, then there has to be recourse to thinking" (Duncker 1945, p. 1).

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 120–137, 2019.
https://doi.org/10.1007/978-3-030-32722-4_8

Given that any situation that involves thought processes can be considered a problem, solving or attempting to solve problems is the typical and, hence, the general function of thought.

Insight problems (problems that require a change in representation by a restructuring process) in particular, can teach us the psychology of thought when we consider the psychological principles that are involved when people are in difficulty. Many different difficulties can be recognized, and they go from self-imposed solution constraints to functional fixedness, to solution mechanization and misunderstanding. We explore the analysis of what inhibits most people from finding the correct solution - or any solution - and from solving creative problems. Themes that strongly emerge include the relative roles of conscious and unconscious processes, the relationship between language and thought, and the roles of special processes particular to insight as against routine processes found widely also in procedural problem solving.

The particularity of insight problems is that they contain one or more critical points susceptible to incompatible representations (Wertheimer 1985) or interpretations (Mosconi 1990, 2016) with the solution, thus leading to an impasse. The representation of the problem must be restructured to allow new research directions. The new interpretation enables to understand the relevant relationships between the data of the problem and allows to find the solution.

Such problems look deceptively simple (Metcalfe 1986) in that they contain little in the way of data, and the number of operations required to arrive at the solution appears limited. However, they are not simple at all. Sometimes, taken in by this apparent simplicity, we are tempted by the first answer that comes to mind, but it is almost always wrong and from that point on, we are faced with an impasse. Sometimes an impasse is encountered immediately, we have no idea how to go about finding the solution, and the problem itself may initially appear impossible to solve (see Bagassi and Macchi 2016; Macchi and Bagassi 2015). These problems may seem bizarre or constructed to provide an intellectual divertissement, but they are a paradigmatic case of human creativity in which intelligence is at its acme. Their study provides a privileged route to understanding the processes underlying creative thought, scientific discovery and innovation, and all situations in which the mind has to face something in a new way, in fact, insight problems have traditionally been considered tests of giftedness (see, for instance, Sternberg and Davidson 1986). An emblematic example of insight problem is the Nine Dots Problem (see Fig. 1).

Nine Dots Problem: Cover these nine dots with four straight lines without removing the hand from the paper.

The wrong attempt is to try to solve the problem by staying within the virtual square. It is necessary to re-interpret the received message to reach the solution (Mosconi 1997), which lies in the observation that it is permissible to cross square boundaries.

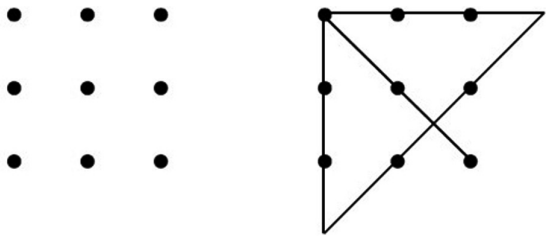


Fig. 1. Nine Dots Problem and its solution.

Insight problems differ from another category of tasks called non-insight problems, or procedural problems, which difficulty lies in the calculations to be made, in the number of operations to be performed and in the amount of data to be processed and remembered. In this case, the solution cannot be reached at once but requires a serial process, step by step, with a gradual simplification of the problem. A well-known example of non-insight problems is the Cryptoarithmetic Problem (Bartlett 1958; Simon and Newell 1971).

Cryptoarithmetic Problem:

$$\begin{array}{r} \text{DONALD} + \\ \text{GERALD} = \\ \hline \text{ROBERT} \end{array}$$

The words DONALD, GERALD, and ROBERT represent three six-digit numbers. Each letter is to be replaced by a distinct digit. This replacement must lead to a correct sum, knowing that:

(a) D = 5;
(b) Each letter represents a unique digit from 0 to 9.

The solver respecting the imposed constraints and starting from a sure data obtains another sure data and proceeds to successive substitutions on the basis of what he has already established, until solving the problem (Mosconi 1997). Simon and Newell (1971) identify in the resolution process a heuristic which involves dealing first with those columns that have the greatest constraints (otherwise there would be 362,880 ways to assign 9 digits to 9 letters): if two digits in the same column are already known, the third can be found by applying the ordinary arithmetic rules.

3 Main Theories of Insight Problem Solving

Different hypotheses have been proposed in the literature to explain the phenomenon of insight, and they can be divided into two main categories: *conscious-work hypotheses* and *unconscious-work hypotheses.*

The cognitivist approach argues that insight problems are solved through a conscious step by step process (Fleck and Weisberg 2004, 2013). In fact, it would be a fully conscious process that takes place in a procedural and stepwise manner. This approach, known as *business-as-usual* (i.e., Bowden et al. 2005), considers restructuring a gradual process in which the solution is processed in a serial way through conscious, reflective thought. The solution is reached at the end of a path that provides a gradual simplification of the problem (Weisberg 2006, 2015) and requires a great working memory capacity. This consolidated theoretical tradition started with Simon and Newell's *Human Information Processing Theory* (Simon and Newell 1971; Newell and Simon 1972) and continued through to the present day (Weisberg 2015). According to these authors, the labyrinth is an appropriate abstract model for human reasoning. A person faced with a problem moves within the problem space just as in a

labyrinth; he searches for the right path, retraces his steps when he comes up against a dead end, and sometimes even returns to the starting point; he will form and apply a strategy of sorts, doing a selective search in the problem space. The labyrinth model, which was devised for procedural problems (e.g., the Cryptoarithmetic Problem or the Missionaries and Cannibals Problem), is also advocated when a change of representation is necessary, extending the selective search process to the meta-level of possible problem spaces for an alternative representation of the problem (e.g., the Mutilated Checkerboard Problem, Kaplan and Simon 1990). Hence, this consolidated theoretical tradition maintains that conscious analytical thought can reorganize data if the initial representation does not work, extracting information from the failure to search for a new strategy (Fleck and Weisberg 2013; Kaplan and Simon 1990; Perkins 1981; Weisberg 2015). According to Weisberg, analytic thinking, given its dynamic nature, can produce a novel outcome; in problem solving in particular, it can generate a complex interaction between the possible solutions and the situation, such that new information constantly emerges, resulting in novelty. It could be said that, when we change strategy, we select a better route; this change is known as "restructuration without insight," which remains on the conscious, explicit layer. Weisberg, rather optimistically, claims that, when the restructuring does not occur and the subject is at an impasse, this "may result in coming to mind of a new representation of the problem. […] That new representation may bring to mind at the very least a new way of approaching the problem and if the problem is relatively simple, may bring with it a quick and smooth solution" (2015, p. 34). However, this model does not explain how this change of representation takes place.

Other hypotheses, classified as *unconscious-work hypotheses*, are inspired by Gestalt psychologists (Dunker 1945; Koffka 1935; Köhler 1925; Wertheimer 1945), which introduced the idea of "productive thought" as the process of elaboration and resolution of insight problem. The term *insight* has in fact been introduced by the Gestaltists to define an intelligent solution process that creates, "produces" the new, distinguishing it from a solution achieved by chance, or based on ideas and behaviors already experienced ("re-productive thinking"). Productive though is characterized by a switch of direction that occurs together with the change in the understanding of an essential relationship between the elements of the problem. Reaching the solution of insight problems (by restructuring) is accompanied, in the solvers' experience, by a manifestation of satisfying surprise, which is called "Aha experience".

The Gestalt vision has been invoked, with different inflections, by the *special process* view of insight problem solving, investigating the qualitatively diverse processes that elude the control of consciousness through spreading activation of unconscious knowledge (Ohlsson 2011; Öllinger et al. 2008; Schooler et al. 1993). This search goes beyond the boundaries of the working memory. The process that leads to the discovery of the solution through restructuring is mainly unconscious, characterized by a period of incubation, and can only be described a posteriori (Gilhooly et al. 2010). The characteristic unconsciousness with which these processes are performed has led to them being defined as automatic, spontaneous associations by chance (Ash and Wiley 2006; Schooler et al. 1993).

In Fig. 2, according to the *special process* view, the phases of insight problem solving are showed (e.g., Ash and Wiley 2006). In the representation phase, the external problem is translated into a mental problem representation. During the solution phase, individuals navigate strategically through the faulty problem space. No solutions can be found consciously within it because in insight problems individuals need to go beyond this initial representation to find the solution. When the possible moves within the problem space are over, the conscious search of the solution is stalled (impasse). During the restructuring phase, individuals could see the problem in a new way through associative processes. If the answer is correct, individuals will exhibit the "Aha experience", otherwise they return in an impasse.

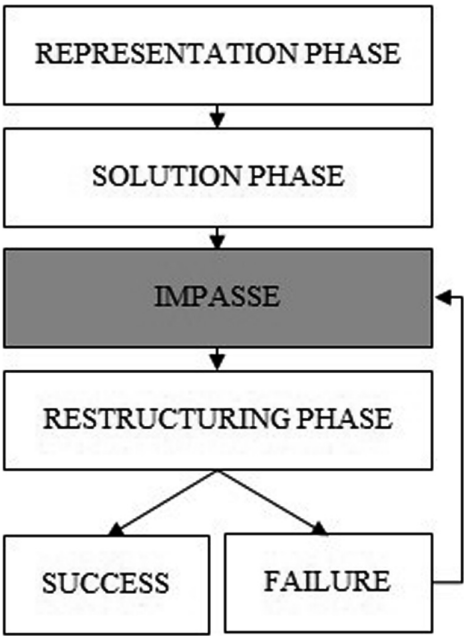


Fig. 2. Phases of insight problem solving. Adapted from Ash and Wiley (2006) and DeCaro et al. (2016)

Both of these views have critical aspects. The *business-as-usual* approach has the limit of blurring the specificity of insight problem together with problems that can be solved with explicit analytical thought processes, thus failing in explaining the creative thinking that derives only from insight problems. In our view, this type of solution seems to result from an unconscious mind-wandering process; then it cannot be attributed to only a reflective and conscious thinking (Baird et al. 2012; Macchi and Bagassi 2012; Smallwood and Schooler 2006). The *special process* view, instead, even

if accounts for the specificity of insight, considering unconscious processes, still view them as merely associative and automatic, contributing to finding the solution almost by chance (Bagassi and Macchi 2016).

According to our perspective, insight problem solving requires of necessity not conscious, implicit processing - incubation - and for this reason, it is a challenge to the current theories of thought, regarding the central role of consciousness, given that, in these cases, restructuring (or change in representation) is not performed consciously. In everyday life too, we sometimes come up against an impasse; our analytical thought seems to fail us, and we cannot see our way out of the dead-end into which we have inadvertently strayed. We know we are in difficulty, but we have no idea what to do next. We have no strategy, and this time failure does not come to our aid with any information that could help us forge forward. In other words, we have come up against a deep impasse. These are situations that change our life radically if we do find a solution: the situation, which a moment before seemed to lead nowhere, suddenly takes on a different meaning and is transformed into something completely new. The same happens in insight problem solving. In our view, in fact, the insight problem solution is achieved by an *unconscious analytic thought* (Macchi and Bagassi 2012, 2015; Bagassi and Macchi 2016), which is a productive, creative thinking, resulting mainly from a covert and unconscious thinking process, and an overall spreading activation of knowledge, which includes a relevant, analytic, goal-oriented search, which goes beyond associative or reproductive thinking. In our perspective, since, insight problems arise from a problem in communication (the so-called *misunderstanding*), and the impasse is due to the failure of the default interpretation, the restructuring process is to be understood, as a form of re-interpretation of the problem, that includes both implicit and explicit processing. Restructuring cannot be fully supported by explicit thinking, which is occupied by fixation and by the default interpretation. The impasse would activate an increasingly focused implicit search leading to a re-interpretation of the data available related to the request.

4 Insight Problem Solving and Working Memory Capacity

Working Memory (WM) could be considered the tip of the balance in favor of *business-as-usual view* or *special process view*. It can be defined as an active multi-component system (Baddeley and Hitch 1974; Baddeley 2000, 2003), composed of an episodic buffer and two slave systems (the phonological loop and the visuospatial sketch pad), controlled by the central executive. WM is often described as a system for storage and processing information in the service of an ongoing task, and at the same time, it allows to focus attention and blocks irrelevant information (Kane and Engle 2003). WM is a system with limited capacity. Individual differences in Working Memory Capacity (WMC) explain the impaired performances in a variety of complex cognitive activities. WMC plays a crucial role in high-level cognitive processes, for example, in reading comprehension, judgment and incremental problem solving (for a review: Barrett et al. 2004).

There is not a universal conceptual definition of WMC, due to the dissent on the mechanisms responsible for the individual differences in WMC[1]. However, there is an operational definition of WMC, used in a large body of research on problem solving, that indicates WMC as the number of items recalled during a complex span task (Redick et al. 2012; Conway et al. 2005). An example of complex span task is the Automated Reading Span Task (aRspan, Redick et al. 2012). In this task, participants have to memorize a letter, while they have to judge if a sentence makes sense or not. After a sequence of sentences and letters (3–7 sequences), participants are asked to recall the letters in order. In contrast to simple span task, which measures only short-term storage capacity, complex span tasks have been proposed to measure short-term maintenance and selective retrieval. According to Unsworth and Engle (2007), higher scores in the complex span task denote greater levels of attentional control.

In problem solving literature, WM processes have been often associated with the conscious experience (e.g., Andrade 2001; Baddeley 1992; Jacobs and Silvanto 2015) and the individual difference in WMC have been made correspond to the various capacity to allocate the resources in executive attention.

In well-defined "analytic" or "incremental" problems, which require a step by step or progressive sub-goals to reach the solution, with attention-demanding processing, the solution came as a result of conscious sequences of step-by-step, specific, mechanistic mental operations.

As already mentioned, according to the *business-as-usual view* (e.g., Bowden et al. 2005), insight and incremental problem are solved through the same underlying mental processes. The only difference is about the achievement of the solution: all-or-nothing in insight problem solving and gradual in incremental problem solving (Weisberg and Alba 1981). In contrast, the *special process* view (Ohlsson 2011; Öllinger et al. 2008; Schooler et al. 1993) argues that insight problems differ from incremental problems in their underlying solution processes because the solution comes as the result of unconscious, associative processes. Therefore, in the first account, a higher WMC should have a positive influence on both types of problems, while in the second view it should have a positive influence on incremental problems solving but not significant influence in insight.

In literature, the relationship between incremental problems and WMC is clear: a higher WMC improves performance in incremental problem solving (e.g., Fleck 2008). However, what happens in insight problem solving is the focus of an interesting debate, because of contradictory perspectives and results (e.g., for significative positive correlations see: Chein et al. 2010; Chronicle et al. 2004; Chuderski and Jastrzębski 2018; for no correlations see Ash and Wiley 2006; Fleck 2008; Gilhooly and Murphy 2005).

In the *special process* view, there is an alternative way: higher WMC can hinder insight (DeCaro et al. 2016), because more focused attention can inhibit the solver from building different representations of the problem from the faulty initial one. They used the *Matchstick Arithmetic Task* (Knoblich et al. 1999), composed of both incremental and insight problems (in Fig. 3 an example).

[1] The main theoretical views are the executive attention view (e.g. Engle 2002), the binding hypothesis (e.g., Oberauer 2009) and the primary and secondary memory view (Unsworth and Engle 2007).

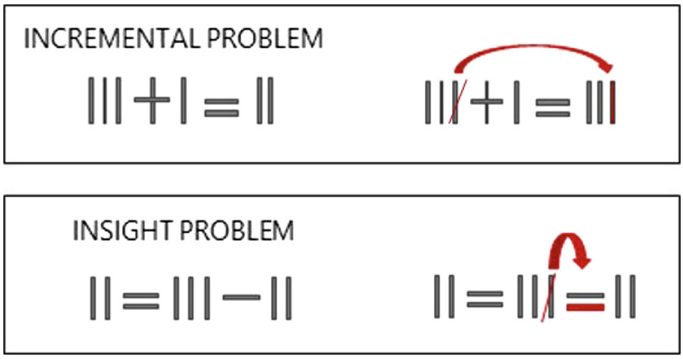


Fig. 3. An example of incremental and insight problems in the *Matchstick Arithmetic Task* (Knoblich et al. 1999), with the solutions.

The authors argued that, on one side, a higher WMC allows a better selection of relevant information and fast construction of the initial problem representation, but on the other side, it can impede the abandonment of the faulty initial problem representation (e.g., Beilock and DeCaro 2007). According to the authors, the initial representation of incremental and insight type of matchstick problems is the same (for example: "I can move the sticks which are part of the numbers, but not those which are part of the signs", or "the sticks could not change orientation"). The initial representation is composed of implicit constraints, which do not influence the solution in an incremental problem but interfere in insight problem solving. A too strong focus on the initial faulty representation due to a high attention control may affect insight problem solving negatively (Wiley and Jarosz 2012) because the solution is not included in that problem space.

Therefore, contrarily to the others, DeCaro et al. (2016) demonstrate that WMC can influence in a negative way the insight. Chuderski and Jastrzębski (2017) tried to replicate these results, with the *Matchstick Arithmetic Task* and other insight problems, without success. On the contrary, they found strong positive correlations between WMC and accuracy in insight problems. These evidences started a debate between the authors for understanding these contradictory findings through the consideration of various kind of individual and situational factors (e.g. DeCaro et al. 2017).

Despite the conflicting results, the idea that WMC can negatively influence insight is very interesting. In our opinion, these evidences do not allow to discriminate between the influence of WMC in every stage of insight problem solving, not only between solution and restructuring phases (DeCaro et al. 2016). Individuals with a higher WMC should be quicker to form an initial problem representation (Jones 2003), because of a more developed capacity in reading comprehension and attention control. However, the initial problem representation itself could be different between individuals with higher and lower WMC. In a recent study (presented at the AIP 2018 Conference, Madrid), Cucchiarini and Macchi showed that the instructions of the matchstick problem could add further constraints to the problem representation, which could increase fixity in subjects with a higher WMC. Using this text, they found that

higher verbal WMC influences in a negative way the accuracy in insight problems, confirming DeCaro et al.'s results. However, removing the additional implicit constraints transmitted by the instructions, without modifying the difficulty and the overall sense of the problems, the effect disappears. When the elements that create fixity much more in individuals with a higher WMC are canceled, the individual differences in WMC do not affect the solution anymore.

These results reflect the complexity of studying the effect of working memory on every phase separately. The phase that could characterize insight as a "special process", distinct from the step by step conscious process of incremental problems, is, in fact, the one that encompasses the restructuring of the problem. Our findings show that WMC, traditionally linked to conscious processes, could not influence insight when the distinction between all phases is considered.

5 The Role of Incubation in Insight Problem Solving: Different Perspectives

To clarify insight problems, it, therefore, seems necessary to explore the relationship between conscious and unconscious, and this can be studied through the phenomenon of incubation.

In the study of creative problem solving it has often been argued that after repeated wrong attempts to solve a problem, suspending the search for the solution, for a certain period of time, can lead to the spontaneous processing of new ideas suitable for the solution (Gilhooly et al. 2015; Schooler et al. 2011). This temporary detachment from the problem with a break in the attentive activity devoted to solving a problem has been first defined by Wallas (1926) "Incubation Period".

To investigate the effects of incubation, the classic laboratory paradigm, called the *Delayed Incubation Paradigm* (Gilhooly 2016), has often been used. In the incubation condition, the participants work on a problem, usually of an insight type, for a certain period, called *preparation* period, after which they are assigned to another task, or activity, to be performed for another predetermined period, defined *incubation* period. Finally, during the *post-incubation* period, participants return to work on the first problem, which had been left pending. To verify if the incubation had effect, or not, the performance in the incubation condition is compared with that of the control condition in which participants continuously work on the insight problem for an amount of time equal to the sum of preparation and post-incubation time of the experimental group (Gilhooly et al. 2015; Segal 2004).

Different incubation tasks can be adopted during the incubation period, and they mostly differ in the degree of cognitive effort required; we can identify three types of tasks often used in incubation studies (respectively from the most effortful to the less): *high cognitive demanding task, low cognitive demanding task,* and *non-demanding task.* Tasks with high cognitive demands (e.g., mental rotation, countdown, memory tests) are aimed at fully occupying the individual's mind to prevent conscious elaboration; while those with low cognitive requirements, such as reading, drawing, do not require focusing all the conscious attention on the task undertaken during the incubation.

Incubation, which remains the core issue in the renewed interest in insight problems is still a crucial question to be solved (Bagassi and Macchi 2016; Fleck and Weisberg 2013; Gilhooly et al. 2013; Gilhooly et al. 2015; Macchi and Bagassi 2012; Sio and Ormerod 2009). A heterogeneous complex of unresolved critical issues underlies the research on this subject (for a review, see Sio and Ormerod 2009) and still revolves around the controversy of the relationship between the conscious and unconscious layers of thought in the solution of insight problems. However, the various mechanisms that have been proposed only describe the characteristics of the solution but do not explain the processes of reasoning that have made the solution possible (these include, for example, eliciting new information, selective forgetting, strategy switching, and relaxing self-imposed inappropriate constraints).

The *selective forgetting hypothesis* (Simon 1966; Smith and Blankenship 1991), a developed version of Woodworth's hypothesis (1938), claims that to allow a "fresh look" to the problem, shifts of attention away from the problem should weaken the activation of the irrelevant concepts that fixate problem solvers' minds, blocking the resolution process. Irrelevant material decays in working memory during the incubation period, and long-term memory accumulates more substantial information, while according to the *fatigue-dissipation hypothesis*, the break simply allows the solver to rest (Seifert et al. 1995).

According to Segal (2004), who introduces the *attention-withdrawal hypothesis,* no process takes place during incubation, which has the only function to divert the attention from the problem, releasing from the false organizing assumption. Following what has been claimed by Gestalt psychologists, Segal states that participants tend to fixate on a false assumption when they try to solve an insight problem, but according to his point of view after encountering the impasse, individuals spontaneously tend to divert attention from the problem. An organizing assumption is necessary to allow individuals to have a mental representation of the problem because it connects all the elements of the problem allowing the solver to understand the problem and act on it. However, when this assumption is false, the achievement of the solution is impossible within the limits of the problem space, since the latter is defective. To solve the problem, therefore, the withdrawal of attention governed by the false assumption must take place and, in the meantime, a certain level of activation of the elements of the problem is needed. Once these elements are no longer bound by the false assumption, the restructuring of the elements of the problem in a configuration suitable for its resolution is favored; thus the solver can apply a different structure, governed by another assumption that leads to the correct solution of the problem. At this point Segal introduces the *returning-act hypothesis*, which follows the *attention-withdrawal hypothesis,* claiming that this mental condition occurs when individuals return to work on the problem after the incubation period. The likelihood of re-adopting the false assumption after incubation is low since it did not work previously. The solver will then tend to apply the correct organizing assumption to the elements of the problem to construct a complete structure that will allow him to solve the problem. It still remains unexplained what leads the problem solver to the new organization.

The incubation effect in insight problem solving is surrounded by uncertainties because different studies have reported different and often contrasting results. As highlighted by Sio and Ormerod (2009), in their meta-analysis, evidence emerged from

experimental studies did not fully support either the *conscious-work hypotheses* or the *unconscious-work hypotheses* (and the following ones). Moreover (as reported by Segal 2004), the number of studies that confirmed the facilitating role of incubation (i.e., Smith and Blankenship 1989), is overall the same as those that have not reported any effect (i.e., Olton amd Johnson 1976).

Sio and Ormerod (2009) proposed an explanation that could account for these conflicting results. According to these authors, who confirm the existence of a positive effect of incubation mainly for a certain class of problems called *creative problems*, there are *procedural moderators* which influence the problem solving processes during the incubation period. A first potential moderator identified by the two authors (Sio and Ormerod 2009) concerns what they call the *nature of the problem*. The various studies taken into consideration by the meta-analysis have applied the incubation to the resolution of different types of problems. Some of them have used the so-called *creative problems*, which require the production of new ideas (there is no right or wrong answer), other studies have instead used the *insight problems* (characterized by a critical point and by a single correct answer). Another moderator identified by the authors regards the *types of tasks* that are used during the incubation period. As we have already mentioned, these can be divided into high, low, and non-cognitive request tasks. Another possible moderator is the *duration of the incubation period*. Longer incubation periods may allow a greater amount of problem solving activities. However, there is no standard definition of what constitutes a long or short incubation period. Kaplan (1990) suggested that to judge whether an incubation period is of long or short duration, the *preparation period* should also be taken into account. In fact, this variable also influences the effects of incubation, since, during the preparation period individuals collect information to form a representation of the problem and make initial attempts that can lead to the impasse (of fundamental importance in the case of insight problems).

Anyway, according to us, a general characteristic that is common to the literature on insight problems in general, and in particular on the incubation-solution relationship, is the total absence of an analysis of the types of difficulty found in individual insight problems. In other words, what makes a difficult insight problem difficult? What kinds of difficulty are we facing? If it were possible to lay them out as a continuum in ascending order of difficulty, we would see that, in fact, the difficulty correlates with the incubation, and this, in turn, with the possible discovery of the solution, thus allowing the restructuring process to occur.

6 Unconscious Analytic Thought

Incubation may offer a measure of the degree and type of difficulty of the problem, as it may vary in length, depending on the degree of gravity of the state of impasse (Macchi and Bagassi 2012; Segal 2004). At this point, the question is what kind of unconscious intelligent thought operates during the incubation to solve these special problems. Through experiments in brain imaging, it is now possible to identify certain regions of the brain that contribute both to unconscious intuition and to the processing that follows. Jung-Beeman et al. (2004) found that creative intuition is the culmination of a series of transitional cerebral states that operate in different sites such as the anterior cingulate of

the prefrontal cortex and the temporal cortex of both hemispheres and for different lengths of time. According to these authors, creative intuition is a delicate mental balancing act that requires periods of concentration from the brain but also moments in which the mind wanders and retraces its steps, in particular during the incubation period or when it comes up against a dead end. According to the hypothesis of the *unconscious analytical thought* (Macchi and Bagassi 2012, 2015; Bagassi and Macchi 2016), incubation should consist in an activity that engages individuals at the conscious level, distancing their explicit attention from the insight problem, but which instead leaves at the unconscious level free cognitive resources, sufficient to process the insight problem. In this way, the insight problem can be processed at the unconscious level during incubation and get rapidly solved after incubation, since the unconscious process is not merely associative (by chance), but is an analytic process relevance addressed.

Incubation is actually the necessary but not sufficient condition to reach the solution. It allows the process but does not guarantee success; however, if it is inhibited, for example, by compelling participants to verbalize, the solution process will be impeded. The study of the verbalization effect, indeed, offers a promising line of research to study the thought processes underlying the solution. In a recent study (Macchi and Bagassi 2012), the "verbalization" procedure was adopted as an indirect method of investigating the kind of reasoning involved in two classical insight problems, the Square and Parallelogram (Wertheimer 1925, see Fig. 4) and the Pigs in a Pen (Schooler et al. 1993, see Fig. 5).

Square and Parallelogram Problem: Given that AB = a, and AG = b, find the sum of the areas of square ABCD and parallelogram EBGD.

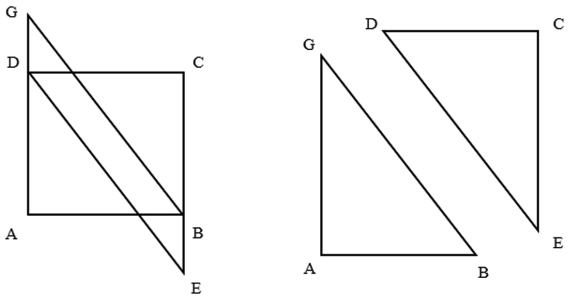


Fig. 4. Square and Parallelogram Problem and its solution.

Pigs in a Pen Problem: Nine pigs are kept in a square pen. Build two more square enclosures that would put each pig in a pen by itself.

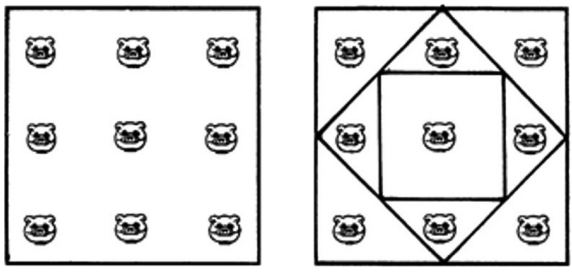


Fig. 5. Pigs in a Pen Problem and its solution.

The investigation focused on whether *concurrent serial verbalization* would disrupt insight problem solving. The hypothesis was that correct solutions would be impaired if a serial verbalization procedure were to be adopted, as it would interfere with unconscious processing (incubation). We found that the percentage of participants who successfully solved insight problems when verbalizing the

process used to reach the solution was inferior to that of the control subjects who were not instructed to do so. In the Square and Parallelogram Problem, there were 80% correct responses in the no-verbalization condition versus only 37% correct responses in the verbalization condition. The difference increased in the Pigs in a Pen Problem: the percentage of correct responses was 12% in the verbalization condition and 87% in the no-verbalization condition.

Our hypothesis was further confirmed by a study on the Mutilated Checkerboard problem (Bagassi et al. 2015), in which the no-verbalization condition significantly increased the number of solutions with respect to the control condition of verbalization, this latter being in accordance with the procedure adopted by Kaplan and Simon (1990). Schooler et al. (1993) also investigated the effect of verbalization on insight, suggesting that language can impair solution and therefore thinking. They claim that "insight involves processes that are distinct from language" (p. 180), given the nonverbal characteristic of perceptual processes. This view follows the traditional dichotomous theory, according to which language, considered extraneous to the higher-order cognitive processes involved in the solution, impairs thinking. We take a different view than Schooler et al. (1993), although their study was extremely stimulating and innovative; in our opinion, language too has a non-reportable side in its implicit, unconscious dimension, which belongs to the common experience and domain. In fact, language as a communicative device is realized by a constant activity of disambiguation, by covert, implicit, unconscious processes, non-reportable in serial verbalization. When the participants in these studies were still exploring ways of discovering a new problem representation, they were not able to express consciously and therefore, to verbalize their attempts to find the solution. Indeed, our data showed that serial "on-line" verbalization, compelling participants to "restless" verbalization, impairs reasoning in insight problem solving; this provides support to the hypothesis of the necessity of an incubation period during which the thinking processes involved are mainly unconscious. During this stage of wide-range searching, the solution still has to be found, and verbalization acts as a constraint, continuously forcing thought back to a conscious, explicit level and maintaining it in the impasse of the default representation. Conscious, explicit reasoning elicited by verbalization clings to the default interpretation, thus impeding the search process, which is mainly unconscious and unreportable.

7 Conclusions

In sum, the main theories in literature on this topic focussed on two contrasting perspectives: the *business-as-usual* view and the *special process* view. For the former, the process underlying the resolution of insight problems is conscious and analytical, while for the latter, it is unconscious and associative. But the characteristic of unconsciousness with which these processes are performed has led to them being defined as automatic, spontaneous associations by the *special process* view (Ash and Wiley 2006; Schooler et al. 1993). On the other hand, the cognitivist approach, known as *business-as-usual*, requires a greater working memory capacity for explaining insight problems, since intelligence identifies with conscious, reflective thought.

Both of these approaches have critical aspects that reveal the complexity of the issue on hand. The *special process* view grasps the specificity of the phenomenon of discovery, which characterizes insight problems and creative thought, but is not in a position to identify the explicative processes because it does not attribute a selective quality to unconscious processes as they continue to be merely associative, automatic, and capable of producing associations that will contribute to finding the solution almost by chance. The limit of the *business-as-usual* approach, on the other hand, is that it levels off the specificity of genuine insight problems, lumping them together with problems that can be solved with explicit analytic thought processes. Finally, this approach makes little progress in explaining the so-called "mysterious event" in the solution of insight problems, relegating them to situations of normal administration that can be dealt with by conscious analytical thought. However, when the solution is mainly the result of a covert and unconscious mind-wandering process, it cannot be attributed to reflective, conscious thinking (Baird et al. 2012; Macchi and Bagassi 2012; Smallwood et al. 2008).

We speculate that the creative act of restructuring implies high-level implicit thought, a sort of *unconscious analytic thought*, informed by relevance, where analytic thought is not to be understood in the sense of a gradual, step-by-step simplification of the difficulties in the given problem, but as the act of grasping the crucial characteristics of its structure. The same data are seen in a different light, and new relations are found by exploring different interpretations, neither by exhaustive searches nor by abstractions but by involving a relationship between the data that is most pertinent to the aim of the task. In this way, each stimulus takes on a different meaning with respect to the other elements and the whole, contributing to a new representation of the problem, to the understanding of which it holistically concurs. Indeed, the original representation of the data changes when a new relation is discovered, giving rise to a gestalt, a different vision of the whole, which has a new meaning. In other words, solving an insight problem - restructuring - means discovering a new perspective, a different sense to the existing relations. The interrelations between the elements of the default interpretation have to be loosened in order to perceive new possibilities, to grasp among many salient cues what is the most pertinent with the aim of the task, or in other words, to reach understanding. This type of process cuts across both conscious and unconscious thought (no more considered as random associative processes). Recent experimental studies and theoretical models are now starting to consider the link between explicit and implicit information processing, in that continuous fluctuation of thought between the focus on the task, the data, and the explicit request, and then the withdrawal into an internal dimension, tempering the processing of the external stimuli to slip into an internal train of thought (stimulus-independent thought - SIT) so as to allow goals other than those that caused the impasse to be considered. This internal processing that takes place during incubation has a neural correlate in the "default mode network" (DMN; Raichle et al. 2001), that is, in a coordinate system of neural activities that continue in the absence of an external stimulus. Moreover, "the sets of major brain networks, and their decompositions into subnetworks, show a close correspondence between the independent analyses of resting and activation brain dynamics" (Smith et al. 2009, p. 13040). It is interesting to note the close correspondence, as neural networks activated, between the brain activity when focused on a task and that recorded during the

resting state. The brain seems to work dynamically in different ways on the same task, at a conscious and not-conscious layer, as the substrate of a restless mind (Smallwood and Schooler 2006; Smallwood et al. 2008). During incubation, when an overall spreading activation of implicit, unconscious knowledge is underway, in the absence of any form of conscious control, relevance constraint allows multilayered thinking to discover the solution, as a result of the restless mind wandering between the implicit and the explicit levels in search of the relationship of the data that would finally offer an exit from the impasse. Rather than abstracting from contextual and more specific elements, as in accordance with the logical approach, we exploit these elements, grasping the gist that provides the maximum of information in view of the aim. Giving sense to something, understanding, does not derive from a summation of semantic units, each with a univocal, conventional meaning. These are simply inputs for that activity of thought, which is crucial to dynamically attribute the most relevant relationship to the recognized aim, in an inferential game that can result in interpretations that can be quite different from those originally intended. Depending on which (new) aim of the task is assumed, the relationship between the items of information changes; when the relationship is changed, the meaning that each element takes on changed too. Incredibly, the result is not cognitive chaos but a sharper understanding.

References

Andrade J (2001) The contribution of working memory to conscious experience. In: Andrade J (ed) Working memory in perspective. Psychology Press, Hove, pp 60–78

Ash IK, Wiley J (2006) The nature of restructuring in insight: an individual-differences approach. Psychon Bull Rev 13:66–73

Baddeley AD (2003) Working memory: looking back and looking forward. Nat Rev Neurosci 4 (10):829–839

Baddeley AD (2000) The episodic buffer: a new component of working memory? Trends Cogn Sci 4(11):417–423

Baddeley AD (1992) Consciousness and working memory. Conscious Cogn 1:3–6

Baddeley AD, Hitch G (1974) Working memory. In: Bower GH (ed) The psychology of learning and motivation: advances in research and theory, vol 8. Academic Press, New York, pp 47–89

Bagassi M, Franchella M, Macchi L (2015) High cognitive abilities or interactional intelligence in insight problem solving? Manuscript under review

Bartlett F (1958) Thinking: an experimental and social study

Bagassi M, Macchi L (2016) The interpretative function and the emergence of unconscious analytic thought. In: Macchi L, Bagassi M, Viale R (eds) Cognitive unconscious and human rationality. MIT Press, Cambridge, pp 43–76

Baird B, Smallwood J, Mrazek MD, Kam JW, Franklin MS, Schooler JW (2012) Inspired by distraction: mind wandering facilitates creative incubation. Psychol Sci 23(10):1117–1122

Barrett LF, Tugade MM, Engle RW (2004) Individual differences in working memory capacity and dual-process theories of the mind. Psychol Bull 130:553–573

Beilock SL, Decaro MS (2007) From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. J Exp Psychol Learn Mem Cogn 33:983–998

Bowden EM, Jung-Beeman M, Fleck J, Kounios J (2005) New approaches to demystifying insight. Trends Cogn Sci 9:322–328

Chein JM, Weisberg RW, Streeter NL, Kwok S (2010) Working memory and insight in the nine-dot problem. Mem Cogn 38:883–892

Chronicle EP, MacGregor JN, Ormerod TC (2004) What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. J Exp Psychol Learn Mem Cogn 30:14–27

Chuderski A, Jastrzębski J (2017) Working memory facilitates insight instead of hindering it: comment on DeCaro, Van Stockum, and Wieth (2016). J Exp Psychol Learn Mem Cogn 43 (12):1993

Chuderski A, Jastrzębski J (2018) Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. J Exp Psychol Gen 147(2):257–281

Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW (2005) Working memory span tasks: a methodological review and user's guide. Psychon Bull Rev 12:769–786

DeCaro MS, Van Stockum CA Jr, Wieth M (2016) When higher working memory capacity hinders insight. J Exp Psychol Learn Mem Cogn 42:39–49

DeCaro MS, Van Stockum CA Jr, Wieth M (2017) The relationship between working memory and insight depends on moderators: reply to Chuderski and Jastrzębski (2017). J Exp Psychol Learn Mem Cogn 43:2005–2010

Duncker K (1945) On problem-solving. Psychological monographs, vol 58, no 270. Springer, Berlin

Engle RW (2002) Working memory capacity as executive attention. Curr Dir Psychol Sci 11:19–23

Fleck JI (2008) Working memory demands in insight versus analytic problem solving. Eur J Cogn Psychol 20:139–176

Fleck JI, Weisberg RW (2004) The use of verbal protocols as data: an analysis of insight in the candle problem. Mem Cogn 32:990–1006

Fleck JI, Weisberg RW (2013) Insight versus analysis: evidence for diverse methods in problem solving. J Cogn Psychol 25(4):436–463

Gilhooly KJ (2016) Incubation in creative thinking. In: Macchi L, Bagassi M, Viale R (eds) Cognitive unconscious and human rationality. MIT Press, Cambridge, pp 301–314

Gilhooly KJ, Murphy P (2005) Differentiating insight from non-insight problems. Think Reason 11:279–302

Gilhooly KJ, Fioratou E, Henretty N (2010) Verbalization and problem solving: insight and spatial factors. Br J Psychol 101(1):81–93

Gilhooly KJ, Georgiou GJ, Devery U (2013) Incubation and creativity: do something different. Think Reason 19:137–149

Gilhooly KJ, Georgiou GJ, Sirota M, Paphiti-Galeano A (2015) Incubation and suppression processes in creative problem solving. Think Reason 21(1):130–146

Jacobs C, Silvanto J (2015) How is working memory content consciously experienced? The 'conscious copy' model of WM introspection. Neurosci Biobehav Rev 55:510–519

Jones G (2003) Testing two cognitive theories of insight. J Exp Psychol Learn Mem Cogn 29:1017–1027

Jung-Beeman M, Bowden EM, Haberman J, Frymiare JL, Arambel-Liu S, Greenblatt R et al (2004) Neural activity when people solve verbal problems with insight. PLoS Biol 2(4):1–11

Kane MJ, Engle RW (2003) Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. J Exp Psychol Gen 132(1):47–70

Kaplan C (1990) Hatching a theory of incubation: does putting a problem aside really help? If so, why? Unpublished doctoral dissertation, Carnegie Mellon University

Kaplan CA, Simon HA (1990) In search of insight. Cogn Psychol 22:374–419

Knoblich G, Ohlsson S, Haider H, Rhenius D (1999) Constraint relaxation and chunk decomposition in insight problem solving. J Exp Psychol Learn Mem Cogn 25:1534–1555
Koffka K (1935) Principles of Gestalt Psychology. Harcourt, Brace and Company, New York
Köhler W (1925) The mentality of apes. Liveright, New York
Macchi L, Bagassi M (2012) Intuitive and analytical processes in insight problem solving: a psycho-rhetorical approach to the study of reasoning. Mind Soc 11(1):53–67 Special issue: Dual process theories of human thought: The debate
Macchi L, Bagassi M (2015) When analytic thought is challenged by a misunderstanding. Think Reason 21(1):147–164
Metcalfe J (1986) Feeling of knowing in memory and problem solving. J Exp Psychol Learn Mem Cogn 12(2): 288–294
Mosconi G (1990) Discorso e pensiero. Il Mulino, Bologna
Mosconi G (1997) Pensiero. In: Legrenzi P (ed) Manuale di psicologia generale. Il Mulino, Bologna, pp 393–453
Mosconi G (2016) A psycho-rhetorical perspective on thought and human rationality. In: Macchi L, Bagassi M, Viale R (eds) Cognitive unconscious and human rationality. MIT Press, Cambridge
Newell A, Simon HA (1972) Human problem solving. Prentice-Hall, Englewood Cliffs
Oberauer K (2009) Design for a working memory. In: Ross BH (ed) The psychology of learning and motivation, vol 51. The psychology of learning and motivation. Elsevier Academic Press, San Diego, pp 45–100
Ohlsson S (2011) Deep learning: how the mind overrides experience. Cambridge University Press, Cambridge
Öllinger M, Jones G, Knoblich G (2008) Investigating the effect of mental set on insight problem solving. Exp Psychol 55(4):269–282
Olton RM, Johnson DM (1976) Mechanism of incubation in creative problem solving. Am J Psychol 89(4):617–630
Perkins D (1981) The mind's best work. Harvard University Press, Cambridge
Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001) A default mode of brain function. Proc Natl Acad Sci USA 98(2):676–682
Redick TS, Broadway JM, Meier ME, Kuriakose PS, Unsworth N, Kane MJ, Engle RW (2012) Measuring working memory capacity with automated complex span tasks. Eur J Psychol Assess 28:164–171
Schooler JW, Ohlsson S, Brooks K (1993) Thoughts beyond words: when language overshadows insight. J Exp Psychol Gen 122(2):166–183
Schooler JW, Smallwood J, Christoff K, Handy TC, Reichle ED, Sayette MA (2011) Meta-awareness, perceptual decoupling and the wandering mind. Trends Cogn Sci 15(7):319–326
Segal E (2004) Incubation in insight problem solving. Creativity Res J 16(1):141–148
Seifert MC, Meyer DE, Davidson N, Patalano AL, Yaniv I (1995) Demystification of cognitive insight: opportunistic assimilation and the prepared-mind perspective. In: Sternberg RJ, Davidson JE (eds) The nature of insight. MIT Press, Cambridge, pp 65–124
Simon HA (1966) Scientific discovery and the psychology of problem solving. In: Colodny R (ed) Mind and cosmos. University of Pittsburgh Press, Pittsburgh, pp 22–40
Simon HA, Newell A (1971) Human problem solving: the state of theory. Am Psychol 21 (2):145–159
Sio UN, Ormerod TC (2009) Does incubation enhance problem solving? A meta-analytic review. Psychol Bull 135(1):94–120
Smallwood J, Schooler JW (2006) The restless mind. Psychol Bull 132:946–958

Smallwood J, McSpadden M, Luus B, Schooler JW (2008) Segmenting the stream of consciousness: the psychological correlates of temporal structures in the time series data of a continuous performance task. Brain Cogn 66:50–56

Smith SM, Blankenship SE (1989) Incubation effects. Bull Psychon Soc 27(4):311–314

Smith SM, Blankenship SE (1991) Incubation and the persistence of fixation in problem solving. Am J Psychol 104(1):61–87

Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, Filippini N, Watkins KE, Toro R, Laird AR, Beckmann CF (2009) Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci USA 106(31):13040–13045

Sternberg RJ, Davidson JE (eds) (1986) Conceptions of giftedness. Cambridge University Press, New York

Unsworth N, Engle RW (2007) The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. Psychol Rev 114:104–132

Wallas G (1926) The art of thought. Jonathan Cape, London

Weisberg RW (2006) Creativity: understanding innovation in problem solving, science, invention, and the arts. Wiley, New York

Weisberg RW (2015) Toward an integrated theory of insight in problem solving. Think Reason 21(1):5–39

Weisberg RW, Alba JW (1981) An examination of the alleged role of "fixation" in the solution of several "insight" problems. J Exp Psychol Gen 110(2):169–192

Wertheimer M (1925) Drei Abhandlungen zur Gestalttheorie. Verlag der Philosophischen Akademie, Erlangen

Wertheimer M (1945) Productive thinking. Harper, New York

Wertheimer M (1985) A gestalt perspective on computer simulations of cognitive processes. Comput Hum Behav 1(1):19–33

Wiley J, Jarosz A (2012) How working memory capacity affects problem solving. Psychol Learn Motiv 56:185–227

Woodworth RS (1938) Experimental psychology. Holt, New York

On Understanding and Modeling in Evo-Devo

An Analysis of the *Polypterus* Model of Phenotypic Plasticity

Rodrigo Lopez-Orellana[1(✉)] and David Cortés-García[2]

[1] Institute of Science and Technology Studies, University of Salamanca, Salamanca, Spain
rodrigo.lopez@usal.es
[2] University of the Basque Country, Leioa, Spain
davidcortesgarcia.dcg@gmail.com

Abstract. In this paper we analyze some particular characteristics of *evo-devo* scientific modeling, starting from a brief analysis of the *Polypterus* model, which is put forward as an explanatory model of the role of developmental plasticity in the evolutionary origin of tetrapods. Evo-devo has brought about an interesting change in the way we understand evolution and it is also posing new challenges for understanding scientific explanation, modeling, experimentation, and the ontological commitments that scientists take on when making theoretical generalizations. Specially in biology, it is necessary to take into account some relevant aspects of modeling that go beyond representation, explanation and the stress on causal relations between phenomena. We approach this type of explanation by accounting for understanding and its relationship to modeling in biology. Thus, our main aim is to elaborate some minimum criteria required to this kind of evo-devo models, so they provide us with effective understanding.

Keywords: Understanding · Models · Explanation · Evo-devo · Tetrapods · Polypterus · Evolution · Phenotypic plasticity · Biology

1 Introduction

As regards the plurality of functions of models in biology, it is worthwhile to ask: What type of explanation can be attempted with the use of models? There are three main types of explanations in biology: *functional* (also teleological[1]), *mechanistic* and *historical-evolutionary*. The most studied ones are functional explanations, due to the problematic character of the notion of *function* [2, p. 13]. Functional explanations are those that do not try to account for material causes but try to explain the relationships between elements. In biology, it is possible to capture this kind of relations by using models.

[1] Traditionally, biological explanations of a functional type, which have the form '*x* serves for *y* in *z*', where *x* is the structure or capacity in question, *y* is the role it plays and *z* is the environmental and historical context in which the organism is located, have been considered as teleological.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 138–152, 2019.
https://doi.org/10.1007/978-3-030-32722-4_9

It is necessary to point out that the notion of function has generated a rich discussion within the philosophy of science. The concept of function has been addressed from different approaches, such us the etiological and dispositional ones, with diverse ontological and epistemological consequences. For this reason, without going into detail, we assume a notion of function simply as a *causal role*. This idea was introduced by Cummins [5, pp. 762–763] and, in a summarized way—following Caponi's analysis [3, pp. 59–60]—, it establishes that '*y* is a function of *x* in process or system *z*, if and only if (1) *x* produces or causes *y*, and (2) *y* has a causal role in the occurrence or the operation of *z*.'[2] We will discuss an application of this definition later in relation to phenotypic plasticity, explained under the analyzed model.

We want to broaden the analysis of model-based explanation by adding the notion of understanding. The problem of understanding has a long tradition in philosophy, and is linked to language pragmatics, ethics, aesthetics and the social sciences. In recent times, it has become of great interest to the philosophy of natural sciences, within the debate around scientific modelling [6, 7, 14]. In the classic distinction within the philosophy of science between *explanation* and *understanding*, the negative psychological character of the second was stressed, thus being relegated to the scope of the social sciences and far from the natural sciences. However, we believe that if the psychological-subjective component of the concept is restricted, as suggested by de Regt [6, 7] and Diéguez [8], it would be possible to achieve an adequate or effective understanding of phenomena. We believe that the notion of understanding has an epistemic relevance for biology, and that it can be illustrated with the use of models in this discipline [15]. In this sense, we call '*effective understanding*'—also '*scientific understanding*'—to the intersubjective understanding of a phenomenon under study that is shared by the scientific community, and that has as its epistemic basis the scientific *experimentation* and *representation* which are associated with the phenomenon. This type of understanding is owed to the interest of the scientist in offering relevant information about the studied phenomena, and to her intention that this information be considered within the general theoretical framework of a given science, which the case of the biologist is evolutionary theory. Models and experiments have a fundamental role in this process.

Our approach and proposal are based on two complementary notions that we believe are related to the model concept. First, we follow Catherine Elgin's definition of scientific understanding [10, p. 327, our italics]:

> […] understanding is a *grasp* of a comprehensive general body of *information* that is grounded in *fact*, is duly responsive to *evidence*, and enables *non-trivial inference*, argument, and perhaps *action* regarding that subject the information pertains to.

[2] Cummins specifies its concept of function as follows: "*x* functions as a Φ in *s* (or: the function of *x* in *s* is to Φ) relative to an analytical account *A* of *s*'s capacity to Ψ just in case *x* is capable of Φ-ing in *s* and *A* appropriately and adequately accounts for *s*'s capacity to Ψ by, in part, appealing to the capacity of *x* to Φ in *s*" [5, p. 762]. Cummins talks about function as 'capacity,' "[the] capacity that a system has to do certain things in a certain way and under certain conditions" [5, pp. 759–761]. Here, we follow Caponi's [3] analysis and characterization.

Elgin's general definition of 'understanding' as "some sort of a cognitive success term" [10, p. 327] captures both cognitive and pragmatic aspects of understanding, since it allows us to properly say something about phenomena, and infer new consequences from them, while allowing us to *act* on the phenomena themselves.

This definition of understanding is closely related to Ian Hacking's [12] notion of *experimentation*, which is the second basic notion of our proposal. Hacking considers experimentation as a dynamic practice, as an *action*, based on the *intervention* and transformation of phenomena, with a certain autonomy with respect to the theory(ies) in question, mainly because the function of the experiments does not reduce to the confirmation or refutation of those theories. Especially in those sciences where the statement of laws is more difficult, uncommon, or problematic—such as biology and social sciences—, the explanatory activity consists mainly in elaborating or using models, which are considered as legitimate and successful explanatory instruments.

For Hacking's experimental realism (experimentalism), reality and experimentation are closely linked; in other words, he asserts that the main function of an experiment is *to create* phenomena. Hacking speaks of creating rather than *discovering*. Given the complexity of nature, it is generally difficult to access the phenomena, as well as to produce them in a stable and controlled way. According to this, we understand why it is common for scientists that their experiments fail: "To ignore this fact is to forget what experimentation is doing […] the real knack of scientists is getting to know when the experiment is working." [12, 230]. In Elgin's terms, this knack reveals the importance of scientific understanding that allows us to act on phenomena and achieve cognitive success.

Thus, "to experiment is to create, produce, refine and stabilize phenomena" [12, pp. 229–230]. It is through this organized and controlled practice that scientists achieve success in their research. But this does not mean that we are assuming a scientific anti-realism about the entities or phenomena of the world. According to Hacking, experimental work provides the best evidence for scientific realism, not only because we can test theories about these entities or phenomena, but rather because we manage to manipulate those entities that cannot be 'observed' and because we manage to produce new phenomena. Thus, we can better understand the different aspects of nature. In our approach, theoretical entities (such as electrons) or the phenomenal relationships assumed by theories (such as phenotypic plasticity, which we will explain below) are seen as *theoretical tools*. They become experimental entities or 'entities for the experimenter' [12, pp. 262–266]. Indeed, manipulation is what allows us to approach an effective understanding of the phenomena of reality. Hacking provides an example:

> The philosopher's favourite theoretical entity is the electron […] electrons have become experimental entities […] In the early stages of our discovery of an entity, we may test the hypothesis that it exists. Even that is not routine. When J.J. Thomson realized in 1897 that what he called 'corpuscles' were boiling off hot cathodes, almost the first thing he did was to measure the mass of these negatively charged particles. He made a crude estimate of e, the charge, and measured e/m. He got m about right, too. Millikan followed up some ideas already under discussion at Thomson's Cavendish Laboratory, and by 1908 had determined the charge of the electron, that is, the probable minimum unit of electric charge. Hence from the very beginning people were less testing the existence of electrons than interacting with them. The more we come to understand some of the causal powers of electrons, the more we can build devices that

> achieve well-understood effects in other parts of nature. By the time that we can use the electron to manipulate other parts of nature in a systematic way, the electron has ceased to be something hypothetical, something inferred. It has ceased to be theoretical and has become experimental. [12, 262]

Of course, it is true that experiments allow us to test hypothesis of the existence of phenomena and, in this way, shield us from falling into a naive realism about entities. But more importantly, as Morrison and Morgan [19] show, models and experiments act as effective *mediators* between phenomena and theories. In fact, following Cartwright, Towfic and Suárez [4, pp. 138–139], we believe that experiments and models are effective instruments for investigation, since they allow us—on one hand—to intervene in the conceptual domain of interpretations and representations of the world, and—on the other hand—to intervene in the phenomenal domain. Within this perspective, both should be considered as tools for scientific understanding.

Specifically, in evo-devo, the biologist's purpose when using a model is to introduce a new dimension—development—in the general evolutionary theory, by considering evolutionary change as an alteration over time of ontogenetic change. We believe that these models are generally used with the intention of obtaining a better understanding of evolutionary diversification phenomena, especially thanks to their integration with evolutionary ecology [22]. In fact, an evo-devo explanation attempts to encompass genetic, epigenetic and historical aspects (the evolutionary history of the lineage in question). In our case, the model of developmental plasticity of tetrapods aims to capture the phenotype-environment interaction from a development perspective. However, we wonder if this model is able to offer an explanation of this phenomenon. Moreover, we can ask ourselves if this model can be correctly characterized as an *explanatory* model and, if so, in what sense, and what kind of explanation it tries to provide. We will show how a model of this kind can account for the local causal structure of an organism's development and the variability of its specific trait. In addition, we will show if it is capable of integrating this variety into a larger model about evolutionary variation and fixation of adaptive novel favourable features on a large scale.

2 The Model of Developmental Plasticity in Tetrapods

In their 2014 paper, Standen, Du and Larsson [23] suggest the following hypothesis:

> *H:* Developmental plasticity, induced by the environment, facilitated the origin of terrestrial traits that led to tetrapods.[3]

The evolution of terrestrial locomotion required the appearance of both anatomical and behavioural traits. The anatomical changes included the appearance of supporting limbs, the decoupling of the pectoral girdle from the skull and the strengthening of the girdle ventrally for support. The predicted behavioural changes include the planting of

[3] The clade Tetrapoda is characterized, mainly, by presenting terrestrial locomotion. Its origin dates back to the Devonian, more than 400 million years ago.

the pectoral fins closer to the midline of the body, thereby raising the anterior body off the ground [23, p. 54]. The specific way by which this "necessary transitions" did occur remains unclear; the main approaches have generally consisted of comparative anatomy (based on the observation of homologies) and in the sketching of phylogenies. Evo-devo approaches, in turn, try to investigate phylogenetic relations by pointing out ontogenetic mechanisms such as phenotypic plasticity,[4] which the authors understand as [23, p. 54]:

> […] the ability of an organism to react to the environment by changing its morphology, behaviour, physiology and biochemistry. Such responses are often beneficial for the survival and fitness of an organism and may facilitate success in novel environments. Phenotypically plastic traits can also eventually become heritable through genetic assimilation […]

Thus, the *Polypterus* experiment is presented as an examination of the developmental plasticity of a sister taxon to the derived group of interest, which, according to the 'flexible stem' model,[5] can be used to appraise ancestral plasticity. The starting premise states that evolutionary transitions, modification and appearance of new traits can be accessible through existing developmental pathways (in this particular case, the *Polypterus* fish), which share certain significant characteristics with the stem clade. In this case, the sister group that is studied is the genus *Polypterus* (also known as 'bichir'), which is the extant fish closest to the common ancestor of *Actinopterygii* and *Sarcopterygii* (Fig. 1). Specifically, the species used in this experiment is *Polypterus senegalus* (Fig. 2a), which resembles to transition fishes as *Tiktaalik* (Fig. 2b) and is postulated by the authors as "one of the best models for examining the role of developmental plasticity during the evolution of stem tetrapods" [23, p. 54].

Polypterus capable of surviving on land and can perform tetrapod-like terrestrial locomotion with its pectoral fins. Nevertheless, it is worth pointing out that *Polypterus* is a predominantly aquatic animal.

At the stage of model manipulation two experimental groups were used: one control group, raised in normal aquatic conditions; and one treatment group, raised in obligatory terrestrial conditions. The aim was to comparatively observe both anatomical and behavioural plastic traits in response to an obligated terrestrial habitat. Terrestrialized fish experienced increased gravitational and frictional forces which were predicted by the scientists to cause changes in the 'effectiveness' of the fishes' locomotory behaviour, as well as changes in the shape of the skeletal structures used in locomotion. The authors also predicted that the plastic responses of the pectoral girdle of terrestrialized *Polypterus* would be similar to the directions of the anatomical changes seen in the stem tetrapod fossil record.

[4] Phenotypic plasticity can be very problematic for a strand of Neo-Darwinian evolutionary thought, the Modern Synthesis, which understands the appearance of new characters as the causal consequence of the random mutations occurrence.

[5] According to West-Eberhard [27, p. 565], the *flexible stem model* of adaptative radiation emphasizes that (1) the origin of variation is an important cause, alongside selection, of adaptative radiation; (2) the phenotypic variants seen in an adaptive radiation often originate within a developmentally flexible ancestral population, or as a result of particular kinds of developmental plasticity present in the common ancestor of the diversified group; and (3) the nature of ancestral developmental plasticity can influence the nature of the radiations.

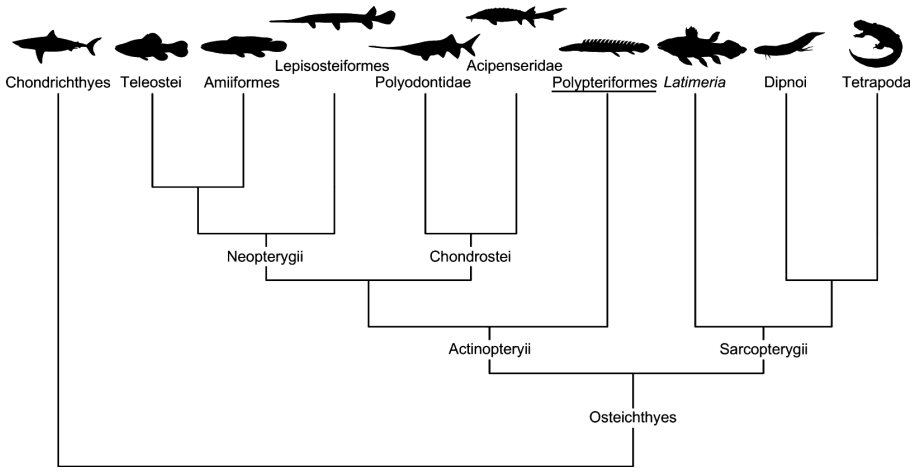


Fig. 1. A phylogenetic tree of the clade Gnathostomata, in which the phylogenetic location of the fish *Polypterus* (Fam. Polypteriformes) is highlighted (underline ours). *Polypterus* is the extant fish closest to the divergence between the classes Actinopterygii and Sarcopterygii. Modified; original from: [28, p. 512].



Fig. 2. a, *Polypterus senegalus*. b, Artistic reconstruction of the sarcopterigian fish *Tiktaalik* (late Devonian). (Pictures via *Wikimedia Commons*).

The study showed that during steady swimming, the fish oscillates its pectoral fins for propulsion, with little body and tail motion (Fig. 3A) while, on land, *Polypterus* walk using a contralateral gait, by using its pectoral fins to raise its head and anterior trunk off the ground and by using its posterior body for forward propulsion (Fig. 3B). Critical differences were observed between terrestrial and aquatic locomotion efficiency. These performance differences suggest that walking is energetically more expensive than swimming [23, pp. 54–55].

Regarding structural changes, the pectoral anatomy of land-raised *Polypterus* exhibits phenotypic plasticity in response to terrestrialization. The clavicle, cleithrum and supracleithrum of the fishes' pectoral girdle create a supporting brace that links the head and the body during locomotion and feeding [11] (Fig. 4a). The clavicle and cleithrum had significantly different shapes in the land-raised and water-raised groups; the treatment group fish had narrower and more elongated clavicles and the horizontal

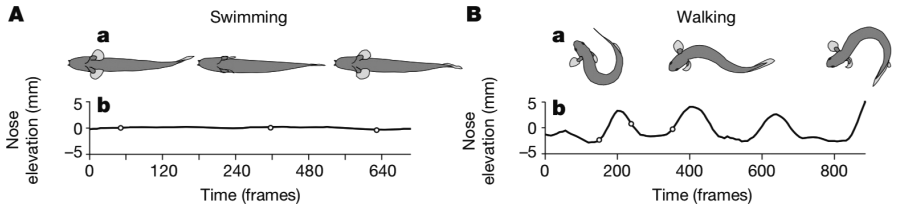


Fig. 3. Kinematic behaviour of swimming (A) and walking (B) *Polypterus*. (Swimming fishes moved farther and faster per fin beat than walking fishes. Moreover, the latter moved their bodies and fins faster, and their nose, tail and fin oscillations were larger. Walking fish also had higher nose elevations, longer stroke durations and greater body curvatures) a, Maximum and minimum body curvature over one stroke cycle. b, Change in nose elevation over several stroke cycles (filmed at 250 frames/s). The circles correspond to the illustrations (from left to right) in a. Modified; original from: [23, p. 55].

arm of their cleitrum also had a narrower lateral surface (Fig. 4b, c). According to the scientists, these skeletal changes—along with other anatomical changes[6]—reflect the need for increased fin mobility in terrestrial environments [23, p. 56].

It is important to note that the causal aspects involved in the experiment can be addressed and determined by the use and manipulation of the statistical technique of multivariate analysis of covariance (MANCOVA). This is a statistical procedure—an extension of the analysis of covariance—that eliminates heterogeneity in the dependent variable by the influence of one or more quantitative variables (covariables) that introduce noise in the cause-effect relation under investigation [13]. Standen, Du and Larsson used this statistical technique in order to determine significant deviations between the two groups (Fig. 4) and to establish a causal relation between those changes and the environmental conditions.

The morphological differences observed between aquatic and terrestrial *Polypterus* bear a remarkable resemblance to the evolutionary changes of stem tetrapod pectoral girdles during the Devonian period. The skeletal changes seen in the treatment group fish are similar to what is observed in stem tetrapods such as *Eusthenopteron* and *Acanthostega* (Fig. 5).

The elongation of the clavicles and the more tightly interlocking cleithrum–clavicle contact, which are common to stem tetrapods and terrestrialized *Polypterus*, might aid in feeding, locomotion and body support in a terrestrial environment. Similar morphologies are also thought to have stabilized the girdle in the earliest tetrapods *Acanthostega* and *Ichthyostega*. Finally, the dissociation of the pectoral girdle from the skull by reduction and loss of the supracleithrum and extrascapular bones allowed the evolution of a neck, an important feature for feeding on land.

[6] When *Polypterus* walks, its fins must move through a larger range of motion than when it swims, forcing the operculum to bend out of the way to accommodate forward fin excursion. This change expands the opercular cavity between the fin and the operculum, providing more space for the pectoral fins to move [23, p. 56].

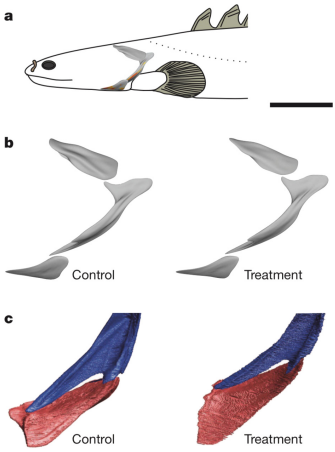


Fig. 4. Anatomical plasticity of Polypterus pectoral girdles. a, Location of the supracleithrum, cleithrum and clavicle in *Polypterus*. Scale bar, 1 cm. b, Left lateral views of the pectoral girdle with the clavicle (bottom), cleithrum (centre) and supracleithrum (top) dissociated for control (left) and treatment (right) group fish. A point-based multivariate analysis of covariance (MANCOVA) with correction for multiple comparisons (false discovery rate estimation) was used to determine the significant differences between the control (n = 7) and treatment (n = 15) *Polypterus* groups. **c**, Close-up anterolateral views of the clavicle (pink)–cleithrum (blue) contact in control (left) and treatment (right) group fish. Modified; original from [23, p. 56].

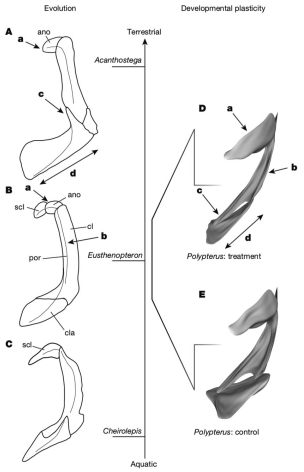


Fig. 5. Scenario for the contribution of developmental plasticity to large-scale evolutionary change in stem tetrapods. Left anterodorsolateral views of the pectoral girdle of: A, B, selected stem tetrapods; **C**, an outgroup; D, land-reared *Polypterus* and E, water-reared *Polypterus*. Comparable developmentally plastic morphologies: a, reduction of the supracleithrum; b, reduction of the posterior opercular chamber edge; c, strengthened clavicle–cleithrum contac and d, narrowing and elongation of the clavicle. ano, anocleithrum; cl, cleithrum; cla, clavicle; por, post opercular ridge (note the ridge in *Cheirolepis* is not distinct but is laterally positioned, as is shown); scl, supracleithrum. Source: [23, p. 57].

On the basis of these observations the authors conclude that [23, p. 57]:

C: the rapid, developmentally plastic response of the skeleton and behaviour of *Polypterus* to a terrestrial environment, and the similarity of this response to skeletal evolution in stem tetrapods is *consistent* with plasticity contributing to large-scale evolutionary change. Similar developmental plasticity in Devonian sarcopterygian fish in response to terrestrial environments may have facilitated the evolution of terrestrial traits during the rise of tetrapods.

The *Polypterus* model is a clear instantiation of evolutionary developmental biology's endeavour to integrate variation, selection and speciation with developmental plasticity in order to explain the origin of physiological, anatomical, biomechanical and behavioural traits. As we pointed out before, and as West-Eberhard argues, this relationship has not yet been satisfactorily explained, and has generated an intense discussion within biology, specifically from the early eighties onwards [27, pp. vii–viii]. To this day, academic scientific circles remain reluctant to accept any extrapolation or generalization of experimental local results on phenotypic plasticity, like *C* is, to a general biological explanation about how organisms have reached their current forms. However, we believe that the model here depicted entails a great advance toward overcoming an explanation of biological evolutionary phenomena exclusively based on the terms of random mutation and selection upon it. The results of the study here presented suggest the eventual introduction of phenotypic plasticity within the paradigm of evolutionary explanation.

3 Developmental Plasticity and Genetic Assimilation

Adopting a broader explanation, Standen, Du and Larsson hold the following hypothesis [23, p. 57]:

> *H':* phenotypic plasticity, as a response to rapid and sustained environmental stresses, may also facilitate macroevolutionary change.

However, they must admit that:

> [m]ulti-generational experiments on terrestrialized *Polypterus* are required to determine the effect of developmental plasticity on the evolution of traits associated with effective terrestrial locomotion.

Despite the warning in this second statement, it is not difficult to set forth criticisms regarding the explanatory gap that sentence *C* entails. It can be argued that scientists undertake an unjustified *explanatory leap* when they perform the extrapolation from ontogenetic conclusions to phylogenetic ones. The evolutionary (or historical) prediction of the origin of tetrapods is supported by a single experiment with a single species. The satisfactory local (ontogenetic) results are taken as evidence for reaching evolutionary (phylogenetic) conclusions, which is unjustified.

For this reason, we think it would be more appropriate to distinguish between two different aspects of the global phenomenon we addressed as phenotypic plasticity: a) *ontogenetic* or *local phenotypic plasticity* (OPP) and b) *phylogenetic* or *large-scale phenotypic plasticity* (PPP). The former accounts for the ability of an organism to perform different phenotypical traits depending on the environmental conditions. The latter is taken as an evolutionary mechanism which consist on the ability of a linage to *adapt* to its environmental conditions through the fixation of multigenerational ontogenetic-type phenotypically plastic traits. We must note that this distinction between OPP and PPP is neither an epistemological nor ontological one, but rather just an analytical one.

The authors try to bridge this explanatory gap by assessing that "phenotypically plastic traits can also eventually become heritable through genetic assimilation" [23, p. 54], which is a necessary condition for achieving an evolutionary explanation, because acquired plastic traits can only have an evolutive value when they become hereditary. However, they assume genetic assimilation *a priori*. The concept of genetic assimilation has been very controversial since Conrad Hal Waddington proposed *canalization* as the mechanism by which genetic assimilation of acquired characters occurs [24–26], in part because the specific genetic and physiological bases of these mechanisms remain unknown even to this day [17]. As Pigliucci, Murren and Schlichting claim, genetic assimilation is regarded as "a process whereby environmentally induced phenotypic variation becomes constitutively produced (i.e. no longer requires the environmental signal for expression)" [21, p. 2362]. It is regarded as the evolutionary outcome of plasticity [21, p. 2366].

According to West-Eberhard, phenotypic evolution occurs as follows [27, p. 140]:

i. *Trait origin*: a mutation or environmental change causes the appearance of a developmental variant expressing a novel trait.
ii. *Phenotypic accommodation* (i.e. a rearrangement of different aspects of the phenotype) to the new trait, made possible by the inherent, pre-existing, plasticity of the developmental system.
iii. *Initial spread of the new variant* facilitated by its recurrence in the population.
iv. *Genetic accommodation* of the novel phenotype, as the result of selection.

More recently, Levis and Pfenning [16, p. 2] clarify that selection can respectively promote either increased environmental sensitivity—which might lead to 'polyphenism'—or decreased environmental sensitivity—in which plasticity is lost and the phenotype becomes canalized (or genetically assimilated).

Summing up, the mechanisms of genetic assimilation[7] would be those that ensured the relationship between the two levels in which phenotypic plasticity can be defined: ontogenetic and phylogenetic ones. Otherwise, the phenotypically plastic traits exhibited by an organism (in our case, the terrestrialized *Polypterus*), or even a population, would not be heritable or, at least, would not be maintained on a time scale broad enough to have an evolutionary relevance.

4 Modeling, Explanation and Understanding in Evo-Devo

With the *Polypterus* model, Standen, Du and Larsson make an *extrapolation* of the local experimental results (phenotypic changes of the fish) that goes from a physiological to a macroevolutionary scale. In other words, they make an estimate or an assumption of the real implication of evolutionary plasticity in the historical origin of terrestrial locomotion, and, consequently, in tetrapods. However, the lack in the description of the mechanisms of genetic assimilation does not guarantee the validity

[7] Genetic assimilation occurs when a trait that was originally triggered by the environment loses this environmental sensitivity (i.e. plasticity) and ultimately becomes 'fixed' or expressed constitutively in a population [9].

extrapolating the local experimental conclusions to the evolutionary conclusions (*C*). Nevertheless, the model offers two significant contributions: (a) from a *functional* explanation, positive local experimental results are presented regarding the existence of OPP in *Polypterus*, which is a first step towards a more general explanation, on the evolutionary plane (*H*); and (b) this model offers an *understanding* of how developmental plasticity could be related to the origin of structural and behavioural traits (*H'*); in other words, it would allow us to understand the *causal role* (function) that plasticity (and other related processes and mechanisms such as genetic assimilation, cryptic variation or canalization) fulfils in the evolutionary system in question. In fact, this is an understanding of how OPP and PPP are related.

The experiment supports an understanding of how novel or stressful environments trigger variation, especially in those organisms that have not been exposed to this environment or that have not undergone major previous adaptations. Thereby, the model encourages us to consider the implication of plasticity in evolution. In other words, plasticity could be a very relevant *functional element* for explanation within the macroevolutionary model, which would enrich a theoretical explanation of large-scale evolution.

Hence, we assert that understanding can be considered as a relevant element to characterizing modeling and the explanatory function of models in biology. We illustrate this idea with the following scheme of effective understanding of a scientific evo-devo model. We are interested in showing that: (1) the experimental results enable a local explanation of the ontogenetic phenotypic plasticity (OPP) in *Polypterus* (Fig. 6 (1)); and (2) an *effective* understanding of the phenomenon is obtained from a global perspective (Fig. 6(2)):

The scheme that we have outlined shows the three instances that encompass an understanding of this type, on which we can make pertinent judgments about the phenomena involved, that is: (i) regarding the phenomena represented by the *Polypterus* model, (ii) regarding the phenomena that are represented with the evolutionary macro-model and (iii) regarding natural history. Next, we suggest four minimum criteria that will make it possible to establish the coherence of this circle of effective understanding. The analysis developed makes the capacity of the *Polypterus* model clear to capture, as an explanatory-functional model, both the complexity of the phenomena it intends to represent and the complexity of the implications of its experimental results and its modeling, in accordance with the hypothesis of the phenotypic plasticity. It is important to note that, as a general feature, evo-devo models have an experimental value in and of themselves. According to this, an evo-devo model manages to offer an effective understanding if:

i. the analogy that is established with the phenomena system represented can be configured empirically, in some way;
ii. the model formulates abstractions that include relevant functional factors that are constitutive of the system and that can be analyzed—in some way—from local causal relationships in the experiment;
iii. it is coherent with the evolutionary theory (or macro-model) of natural selection, and can be integrated and articulated with prior knowledge of biology; and
iv. its predictions contribute to the global understanding of evolutionary history and the theory of evolution, thanks to experimental success.

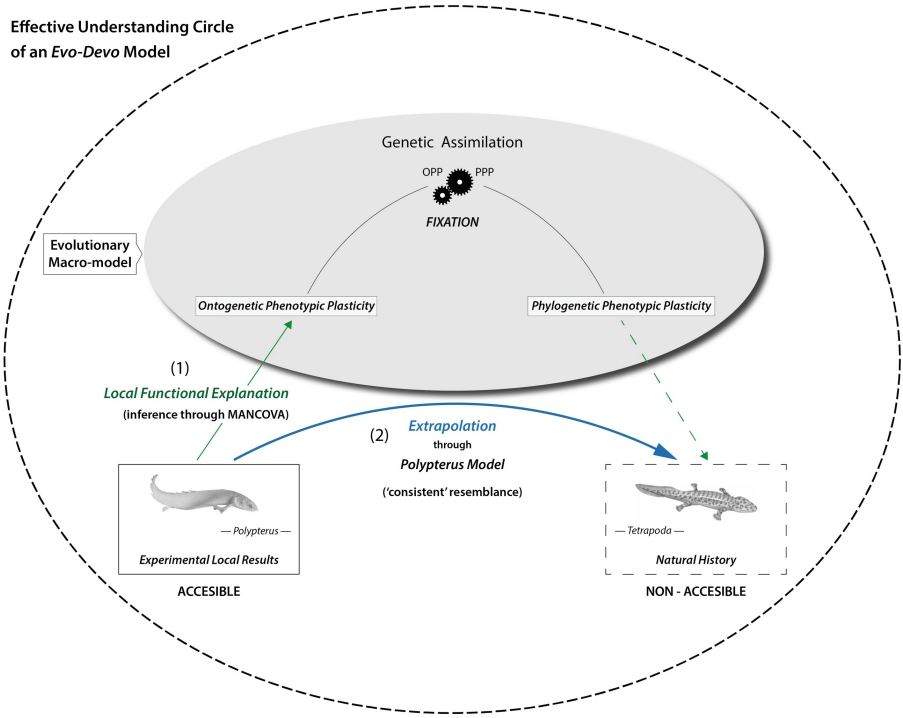


Fig. 6. General scheme of effective understanding of an evo-devo model (*Polypterus*). (1) Local functional explanation of phenotypic plasticity through the analysis of causal relationships given in the experiment; (2) extrapolation of the experimental results through the model.

With respect to points iii. and iv., it is necessary to point out that the evo-devo explanation proposes the inclusion in the global explanation of biology of aspects that had been ignored by the Modern Synthesis, such as the ontogenetic dimension and questions related to form (morphogenesis, body plans), causal embryology and developmental genetics. In this way, we cannot say that evo-devo implies a radical break with the Modern Synthesis. We believe that evo-devo models can be integrated into the theoretical context of the Modern Synthesis (in the sense of the Extended Evolutionary Synthesis), despite representing a break with some of its assumptions. This integration would be guaranteed by (a) the coherence between the explanations given by the model in question and the fundamental theoretical content of the theoretical body of biology itself, and (b) the possibility of establishing the mechanical-functional basis of such explanations in physicochemical, genetic and/or cellular terms.

Finally, the hypothesis of the implication of phenotypic plasticity in the appearance of tetrapod traits would be verified when: (1) the conditions for an effective extrapolation are met by the model, based on the replication of experimental results (2) this knowledge can be integrated into an evolutionary theory, which means that the role of

plasticity in evolution can be elucidated at different scales and in different taxonomic groups; and (3) it is possible to offer a mechanistic explanation in molecular and cellular terms—and other ones—of character fixation and genetic assimilation processes.

5 Conclusions

Developmental systems that evo-devo studies are of a complex organism-environment interaction nature. The *Polypterus* model assumes a homology between these modern fishes and the first tetrapods from a causal basis between common ancestry and shared development mechanisms. It is obviously impossible to provide a set of criteria that allow us to establish with *certainty* this homology relationship—or any other. Therefore, Standen, Du and Larson do the right thing: they establish this homology by appealing to the common origin and, secondarily, to the functional morphology of these animals. As Love [18] suggests, this seems to be the only way to ground an effective homology relationship. This is the reason why we hold that understanding plays an important role for scientific explanation within evo-devo models.

> In order to genuinely encapsulate the epistemological heterogeneity of biology, we need to devote more attention to how biological knowledge is actually structured, as well as rethinking the nature of evolutionary theory in a broader sense than is understood by many philosophical commentators. An overly narrow view of functions and ignorance of diverse conceptual practices in biology should no longer obscure our attempts to elucidate the dynamics of reasoning and explanation in biology [18, p. 706].

In this sense, we propose the circle of effective understanding for an evo-devo model as an effort to characterize the scientific practices of modeling and explanation in biology from a broader and more dynamic perspective. The criteria we provide are not *fixed*, they can be revised and modified, with the intention of better specifying the *effectiveness* of a model for giving us an understanding, and—in this way—better delimiting the attempted explanation.

In this paper, our analysis has been restricted—because of extension and precision issues—to the *Polypterus* model. However, we believe that this analysis can be applied to other cases to configure a general characterization: for instance, for *Lepidobatrachus* model, that tries to elucidate the embryogenesis and morphogenesis of early vertebrate [1], and also models of tooth development, about the origin of hominoid molar diversity [20].

One of the most prominent endeavours of evo-devo is to investigate phylogeny throughout the study of ontogenetic processes. For this purpose, a re-evaluation of the mechanisms of evolution is required; the role of natural selection must be measured and clarified, and the generators of novelty have to be thoroughly described. Because of this reason, twentieth-first century evo-devo brings about a renewed interest in concepts such as morphospaces, developmental and anatomical constraints, plasticity, and so on. Particularly, theoretical and empirical inquiry on plasticity constitute an attempt to link ontogeny and phylogeny.

Finally, we conclude that evo-devo offers an understanding of evolution that is not based exclusively on the neo-Darwinian perspective of the Modern Synthesis, which understands evolutionary change as a consequence of the *intervention* of natural

selection on introduced characters by the occurrence of random mutations. Briefly stated, evo-devo suggests that the Modern Synthesis is not erroneous, but *incomplete*. In the specific case addressed here, phenotypic plasticity is introduced as a local phenomenon that plays a relevant role in the explanation and understanding of macroevolution. Hence, following Cummins' terminology [5], *Polypterus* model tries to show that phenotypic plasticity (local function) plays a *causal role* in the occurrence of evolutionary changes (general function).

References

1. Amin N, Womble M, Ledon-Rettig C, Hull M, Dickinson A, Nascone-Yoder N (2015) Budgett's frog (lepidobatrachus laevis): a new amphibian embryo for developmental biology. Dev. Biol. 405(2):291–303
2. Braillard P-A, Malaterre C (2015) Explanation in biology: an introduction. In: Braillard P-A, Malaterre C (eds) Explanation in biology. An enquiry into the diversity of explanatory patterns in the life sciences, vol 11. Springer, Dordrecht, pp 1–28
3. Caponi G (2010) Análisis funcionales y explicaciones seleccionales en biología. Una crítica de la concepción etiológica del concepto de función. Ideas y Valores 59(143):51–72. https://revistas.unal.edu.co/index.php/idval/article/view/36654/38573
4. Cartwright N, Shomar T, Suárez M (1995) The tool box of science. Tools for the building of models with a superconductivity example. In: Herfel WE, Krajewski W, Niiniluoto I, Wójcicki R (eds) Theories and models in scientific processes. Editions Rodopi B.V., Amsterdam, pp 138–149
5. Cummins R (1975) Functional analysis. J Philos 72(20):741–765. https://doi.org/10.2307/2024640
6. de Regt HW, Dieks D (2005) A contextual approach to scientific understanding. Synthese 144(1):137–170. https://doi.org/10.1007/s11229-005-5000-4
7. de Regt HW, Leonelli S, Eigner K (2009) Focusing on scientific understanding. In: de Regt HW, Leonelli S, Eigner K (eds) Scientific understanding. Philosophical perspectives. University of Pittsburgh Press, Pittsburgh, pp 1–17
8. Diéguez A (2013) La función explicativa de los modelos en biología. Contrastes. Rev Internacional de Filosofía 18:41–54. https://doi.org/10.24310/Contrastescontrastes.v0i0.1157
9. Ehrenreich IMW, Pfennig D (2016) Genetic assimilation: a review of its potential proximate causes and evolutionary consequences. Ann Bot 117(5):769–779. https://doi.org/10.1093/aob/mcv130
10. Elgin CZ (2009) Is understanding factive? In: Haddock A, Millar A, Pritchard D (eds) Epistemic value. Oxford University Press, Oxford, pp 322–330
11. Gosline WA (1977) The structure and function of the dermal pectoral girlde in bony fishes with the particular reference to ostariophysines. J Zool Soc Lond 183:329–338. https://doi.org/10.1111/j.1469-7998.1977.tb04191.x
12. Hacking I (1983) Representing and intervening. Introductory topics in the philosophy of natural science. Cambridge University Press, Cambridge
13. Huberty CJ, Petoskey MD (2000) Multivariate analysis of variance and covariance. In: Tinsley HEA, Brown SD (eds) Handbook of applied multivariate statistics and mathematical modeling. Academic Press, San Diego, pp 183–208

14. Knuuttila T, Merz M (2009) Understanding by modeling an objectual approach. In: de Regt HW, Leonelli S, Eigner K (eds) Scientific understanding. University of Pittsburgh Press, Pittsburgh, pp 146–168
15. Leonelli S (2009) Understanding in biology: the impure nature of biological knowledge. In: de Regt HW, Leonelli S, Eigner K (eds) Scientific Understanding. Philosophical Perspectives. University of Pittsburgh Press, Pittsburgh, pp 189–209
16. Levis NA, Pfennig DW (2019) Phenotypic plasticity, canalization, and the origins of novelty: evidence and mechanisms from amphibians. Semin Cell Dev Biol 88:80–90. https://doi.org/10.1016/j.semcdb.2018.01.012
17. Loison L (2018) Canalization and genetic assimilation: reassessing the radicality of the waddingtonian concept of inheritance of acquired characters. In: Seminars in cell and developmental biology (in press). https://doi.org/10.1016/j.semcdb.2018.05.009
18. Love A (2007) Functional homology and homology of function: biological concepts and philosophical consequences. Biol Philos 22:691–708
19. Morrison M, Morgan MS (1999) Models as mediating instruments. In: Morgan MS, Morrison M (eds) Models as mediators. Perspectives on natural and social science. Cambridge University Press, Cambridge, pp 10–37
20. Ortiz A, Bailey SE, Schwartz GT, Hublin J-J, Skinner MM (2018) Models of tooth development and the origin of hominoid molar diversity. Sci Adv 4(4) (2018). https://doi.org/10.1126/sciadv.aar2334
21. Pigliucci M, Murren CJ, Schlichting CD (2006) Phenotypic plasticity and evolution by genetic assimilation. J Exp Biol 209(12):2362–2367. https://doi.org/10.1242/jeb.02070
22. Santos ME, Berger CS, Refki PN, Khila A (2015) Integrating evo-devo with ecology for a better understanding of phenotypic evolution. Briefings Funct Genomics 14(6):384–395. https://doi.org/10.1093/bfgp/elv003
23. Standen EM, Du TY, Larsson HCE (2014) Developmental plasticity and the origin of tetrapods. Nature 513:54–58. https://doi.org/10.1038/nature13708
24. Waddington CH (1942) Canalization of development and the inheritance of acquired characters. Nature 150:563–565. https://doi.org/10.1038/150563a0
25. Waddington CH (1953) Genetic assimilation of an acquired character. Evolution 7(2):118–126. https://doi.org/10.2307/2405747
26. Waddington CH (1961) Genetic assimilation. Adv Genet 10(C):257–293. https://doi.org/10.1016/s0065-2660(08)60119-4
27. West-Eberhard MJ (2003) Developmental Plasticity and Evolution. Oxford University Press, Oxford
28. Wilhelm BC, Du TY, Standen EM, Larsson HCE (2015) *Polypterus* and the evolution of fish pectoral musculature. J Anat 226(6):511–522. https://doi.org/10.1111/joa.12302



Conjuring Cognitive Structures: Towards a Unified Model of Cognition

Majid D. Beni(✉)

Department of History, Philosophy and Religious Studies,
Nazarbayev University, Astana, Kazakhstan
davoody1980@gmail.com

Abstract. There are different philosophical views on the nature of scientific theories. Although New Mechanistic Philosophy (NMP) and Structural Realism (SR) are not rival theories strictly speaking, they reinterpret scientific theories by using different kinds of models. While NMP employs mechanistic models, SR depends on structural models to explicate the nature of theories and account for scientific representation. The paper demonstrates that different kinds of models used by NMP and SR result in quite different evaluations of the unificatory claims of a promising theory of cognitive neuroscience (the Free Energy theory). The structural realist construal provides a more charitable reading of the unificatory claims of the Free Energy Principle. Therefore, I conclude, it has an edge over NMP in the present context.

1 Introduction

The Prediction Error Minimization theory (PEM, for short) is a promising theory of computational neuroscience. The Free Energy Principle (FEP) provides a grand unifying framework that subsumes PEM. FEP and PEM are viable instances of Bayesian theories of cognition (also perception and action) and life. FEP integrates theoretical devices from cognitive psychology, theoretical biology, information theory, and reinforcement learning, and it comprises insights from Helmoltzian neurophysiology, statistical thermodynamics, and machine learning. The unifying power of the free energy formulation of PEM has been praised by some advocates. For example, Karl Friston submits that PEM can unify cognition, perception, learning, memory, etc., and it can provide the basis for a unified science of cognition (Friston 2010). Andy Clark claims that PEM grounds "a deeply unified theory of perception, cognition, and action" (Clark 2013, 186). And Jakob Hohwy argues that PEM explains *everything* about the mind and its maximal explanatory scope, and it supports a unified approach to mental functions (Hohwy 2013, 242, 2014, 146). There are, however, those who voice scepticism as regards the unifying power of Bayesian cognitive science in general and FEP in particular (Colombo and Hartmann 2015; Colombo and Wright 2016).

For coming to the final version of this paper, I received constructive comments from Marcin Milkowski and two anonymous reviewers of this volume. The debt is gratefully acknowledged. I also thank the editors of this volume sincerely for their collaboration.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 153–172, 2019.
https://doi.org/10.1007/978-3-030-32722-4_10

At the heart of this disagreement lies an intriguing philosophical point concerning the issue of perspectives. One aim of this paper is to shed some light on different perspectives which bear on the evaluation of the unificatory powers of FEP (and PEM). Two perspectives that bear differently on this evaluation are New Mechanistic Philosophy (NMP) and Structural Realism (SR). Roughly speaking, NMP presumes that viable scientific explanations are based on how actually models of the operation of causal mechanisms (Craver 2007; Glennan 2017). I argue that some negative evaluations of the unificatory claims of FEP (e.g., Colombo and Wright 2016, 2018) have their roots in NMP, which invokes mechanistic models to explain cognitive phenomena. What strikes me as odd is that the critics adopted mechanistic models as the right tools for assessing FEP, despite being clear about the inconsistency between FEP and the mechanistic paradigm of explanation (Colombo and Wright 2018, Sect. 4). It is possible though to take a more charitable approach to this issue. The paper shows how some versions of SR (French 2011, 2013, 2014) can be developed to a perspective that accommodates a more charitable understanding of the unificatory claims of FEP. I argue that the form of unification that can be accomplished at the meta-theoretical Bayesian framework of FEP is legitimate from the perspective of SR, which uses structural models to account for scientific representations. We may construct our mechanistic explanations of cognitive phenomena upon our understanding of functional relations between biophysical component parts of organisms, but those explanations would be local. FEP, however, comes with global explanatory ambitions. I argue that it is in virtue of its unifying meta-theoretical formal framework that FEP accommodates global explanation of the fundamental features of life and cognition. To be clear, I do not assume that a unificatory conception of scientific explanation is per se to be preferred over a mechanistic conception of explanation. Nor do I take the supremacy of unificatory explanations for granted. My point is that a mechanistic view on explanation is just too unsympathetic to FEP's unificatory claims, and SR provides a stance for developing a more charitable evaluation of FEP's unificatory power.

The structure of the paper is straightforward. I canvass FEP and PEM. Then I explain that some negative evaluations of the unifying pretences of Bayesian psychology (in general) and FEP (in particular) have their roots in a mechanistic view. I argue that although there is a recognised inconsistency between FEP and the mechanistic view, the critics have adopted an unsympathetic mechanistic approach to the evaluation of the unificatory power of FEP. The remaining sections of the paper flesh out a charitable evaluation of the unificatory powers of FEP. They show how some version of SR bears favourably on the unificatory claims of FEP.

2 FEP and Its Unificatory Scope

The Predictive Error Minimisation theory (PEM) is a successful theory of contemporary computational neuroscience (Alderson-Day et al. 2016; Horga et al. 2014; Rao and Ballard 1999; Seth 2014). According to this theory, the brain first forms internal models and then applies its models to reality. The brain succeeds in providing precise representations of the causal structure of reality by minimising the discrepancy between its generated models and reality. PEM employs a (variational) Bayesian formalism to

model the relevant mechanisms. To be more precise, the free energy formulation of PEM can be understood as a special form of Bayesian theories of cognition. According to this construal, the brain uses Bayesian inferences to reduce the amount of its uncertainty about the hidden causes of perception.

There are explicitly Bayesian brain theories according to which the brain invokes Bayesian models to explain how neurons encode the structure of sensory information in terms of probability distributions (Knill and Pouget 2004; Seth 2015). Such models account for the nature of neural information processing in terms of the conditional probability functions that the brain forms to infer the structure of causes given the sensory information. The relationship between PEM and the Bayesian brain theory is that, according to the free energy formulation of PEM, the brain's mechanisms of minimising prediction error will approximate Bayesian inference over the long-term. FEP affords a certain way of approximating Bayesian inference which subsumes the brain's predictive coding algorithms. It is true that one could be an advocate of PEM and not subscribe to FEP, opting for a different kind of approximation and a different set of formal tools. However, the free energy formulation of PEM allows us to unify the brain's predictive coding, the Bayesian brain theory, and basic mechanisms of adaptation and survival on the basis of the Bayesian formalism (Friston 2010; Friston and Stephan 2007; Friston et al. 2012).

The Free Energy Principle (FEP) specifies basic principles of life and cognition by providing a comprehensive framework that captures the main insights of the Bayesian brain theory, Hebbian plasticity, evolutionary neuroscience, probability theory, information theory, optimal control and game theory (Friston 2010; Ramstead et al. 2017). It is in virtue of its inclusive mathematical formalism that FEP could relate insights from these diverse fields. FEP provides a Bayesian measure of the amount of the brain's uncertainty about hidden causes, which is the same thing as the discrepancy between the brain's internal models and causal structures in the world. When the brain's models of the world do not conform to reality, the brain is surprised. It is desirable to keep the amount of surprise low because the brain has to waste more energy to react to unforeseen (i.e., surprising) accidents. The organism prevents this eventuality by reducing the entropy (or the average surprise) of the system. FEP models this mechanism. Free energy is an information-theoretic measure that bounds the amount of the surprise on sampling data given a generative model (Friston 2010, 1), where surprise or self-information is the negative log probability of an outcome. The theory characterises generative models as probabilistic models that specify "the likelihood of data, given their causes (parameters of a model) and priors on the causes" (Friston 2010, 2–3). Entropy is the average surprise, and free energy puts an upper bound on the entropy of the informational exchange between the organism and its environment.

When applied to cognitive systems, FEP characterises mechanisms of prediction error minimisation. FEP explains how the brain produces probabilistic models of generative predictions and uses sample data to update its models of the world by invoking (variational) Bayesian mechanisms. In this context, perception—or knowledge of the world's causal structure—would be defined as "the process of inverting the likelihood model (mapping from causes to sensations) to access the posterior probability of the causes, given sensory data (mapping from sensations to causes)" (Friston 2010, 3). As this short introduction indicates, articulating PEM in terms of free energy

renders PEM more viable (from a biological point of view) and enhances its unificatory power by grounding the basic mechanisms of cognition (also perception and action) in evolutionary mechanisms of adaptation and survival (see Ramstead et al. 2017). Formulating PEM in terms of FEP makes it possible to elucidate the brain's capacity for representing the causal structure of the world on the basis of evolutionary facts about the tendency of organisms to stay in stable equilibrium with their environment. By minimising the amount of its prediction error or surprise through *active inferences,* the brain maximises the survival of the organism. Organisms can actively infer the structure of the world in the two following ways: "they can change sensory input by acting on the world or they can change their recognition density by changing their internal states" (Friston 2010, 3).

Let us recap. PEM presumes that the brain invokes Bayesian inferences to infer the causal structure of the world. The relationship between PEM and the Bayesian Brain Theory is not straightforward. Strictly speaking, the brain does not form Bayesian inferences, because Bayesian inferences over complex domains (like the world) are computationally intractable (Sanborn and Chater 2016). However, arguably, the brains are capable of variational approximation to probability distributions, presuming that the approximations could capture the intractable computational states that arise in the Bayesian framework. Variational Bayesianism lacks the ideal accuracy with which Bayesian inferences calculate the probability of likelihoods. However, the fact that implementation theories such as PEM rely on variational Bayesianism does not exclude them from the rank of Bayesian theories. Of course, the fact that, generally, scientific models (forged in terms of Bayesianism, variational Bayesianism, set theory, model theory, category theory, etc.,) are refined idealised tools that apply to reality only approximately does not mean that we cannot rely on scientific models to account for the representational capacity of theories (Godfrey-Smith 2009; Weisberg 2013). The same holds true of the brain's reliance on variational Bayesianism. The fact that the brain uses approximate Bayesianism does not need to mean that Bayesian inferences do not provide a framework for explaining the brain's representational capacity.

Finally, I have to point out that the unifying virtue of the free energy formulation of PEM has been praised. It has been claimed that it is "a deeply unified theory of perception, cognition, and action" (Clark 2013, 186). Allegedly the theory explains *everything* about the mind because its maximal explanatory scope underpins a unified approach to mental functions (Hohwy 2013, 242, 2014, 146). The optimism of PEM-theorists about the unificatory powers of PEM is quite understandable. PEM promises to present the most fundamental principles of cognition and life. And fundamental theories of science usually come with a considerable unificatory trajectory. For instance, Newtown's theory of gravitation unifies theories of planetary motion and terrestrial motion, and Maxwell's equations unify theories of magnetism and electricity, etc., (Kitcher 1989). Because FEP unveils the first principles of life and cognition, it is expected to have great unifying power. However, there are critics who challenge these unificatory claims. For example, according to Colombo and Wright (Colombo and Wright 2016), FEP fails to accommodate the promised overarching unificatory framework. This negative evaluation has its roots in a mechanistic view which is allegedly inconsistent with FEP (Colombo and Wright 2018, Sect. 4). I shall unpack this last remark in the two next sections.

3 Mechanistic Explanations

NMP is a remarkable theory of the philosophy of science. It aims to set a paradigm of the mechanical explanation of phenomena (in special sciences) based on the organisation of component parts and their underpinning mechanisms. Explaining a phenomenon consists in the elucidation of its underlying mechanisms. As Bechtel and Abrahamsen (2005, 423) remarked "a mechanism is a structure performing a function in virtue of its component parts, component operations, and their organisation. The orchestrated functioning of the mechanism is responsible for one or more phenomena". Accordingly, the mechanistic account of scientific explanation underlines the role of mechanistic components in producing understanding. Kaplan defines mechanistic explanation in the following way:

> A model of a target phenomenon explains that phenomenon when (a) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism. (Kaplan 2011, 272)

It is worth noting that the mechanistic account of explanation is essentially different from Kitcher's unificationist account of explanation according to which it is best to explain phenomena on the basis of stringent patterns, namely, by using "the same pattern of derivation again and again" and reducing "the number of facts we have to accept as ultimate" (Kitcher 1989, 432). Being restrictive or stringent is not a condition on the plausibility of mechanistic explanations, which may include details about various components that contribute to the elucidation of the function of target mechanisms. This indicates that mechanistic explanations are not committed to unification as an explanatory virtue, and mechanistic explanations could be plausible without being unifying. A clarification regarding mechanistic models' perspective on the notions of unification and integration may shed some light on this claim.

Mechanists' views on the issue of unification are not unanimous. There are mechanists such as James Tabery (Tabery 2014; Tabery et al. 2014) who renounce the goal of unification completely and suggest that NMP lines up with outright explanatory pluralism. There are also mechanists who assume that NMP may allow for some forms of integration, but they point out that mechanistic integration is collaborative and piecemeal, and it is in harmony with explanatory pluralism (Mitchell 2003). Then again, Craver is a mechanist who suggests that mechanistic unification can be "achieved by showing how functional analyses of cognitive capacities can be and in some cases have been integrated with the multilevel mechanistic explanations of neural systems" (Piccinini and Craver 2011, 284). Notions of unification and integration are somewhat used interchangeably in this phrase. However, Milkowski (2016), argues that Craver conflates between notions of 'unification' and 'integration'. According to Milkowski, explanatory unification—defined as "the process of developing general, simple, elegant, and beautiful explanations"—is not the same thing as explanatory integration—which is "the process of combining multiple explanations in a coherent manner" (Miłkowski 2016, 16).

In all, unification can be a desirable feature of mechanistic explanations, but it does not need to be an indispensable virtue of mechanistic explanations (see Glennan 2017, 210–13). This is in line with the point that I fleshed out earlier (with reference to

Kitcher's and Kaplan's works) about the divergence between mechanist and unificationist kinds of explanations. The unificatory account of explanation is supposed to be based on as small a number of patterns as possible. However, a mechanist can contend that there may exist a large number of basic mechanisms in fact, and therefor unification on the basis of the least number of mechanisms cannot be a viable criterion of the plausibility of explanations (Glennan 2002, S352; Miłkowski 2016, 26). Integration, on the other hand, allows for accommodating as large a number of mechanisms as required, and thus mechanists can embrace the notion of integration as an explanatory norm. The take-home point is that NMP does not support a unificationist theory of explanation.

4 Mechanistic Evaluation of the Unifying Power of FEP

In the previous section, I pointed out that mechanistic explanations are essentially different from unificationist explanations. Mechanistic explanations are not concerned with reducing the number of independent facts by subsuming them under an explanatory store (as a set of patterns that unifies explananda). Here, I show that the mechanistic approach to explanation bears unfavourably on the evaluation of the explanatory power of FEP.

Some critics of the unificatory pretences of Bayesian approaches to cognition advocated a mechanistic view on explanation. For example, Colombo and Hartmann (2015) submitted that it may be possible to consider unificatory virtues when providing mechanistic explanations. However, they argue that unifying patterns possess explanatory power only when they identify with causal unifications that contribute to the specification of actual relations between models[1] of mechanisms (Colombo and Hartmann 2015). In other words, abstract unifying patterns should be grounded in actual causal unifying mechanisms to be taken seriously in mechanistic explanations. This is because NMP extols the role of casual mechanistic elements that figure in the explanations over the unifying virtue of explanations. Colombo and Hartmann build upon this fundamental point about the nature of plausible (mechanistic) explanations to submit a negative assessment of the unificatory power of Bayesian theories of cognition. A similar course has been pursued by Colombo and Series.

Colombo and Series' (2012) paper explores the capacity of Bayesian models of cognition for producing mechanistic explanations of cognitive phenomena (such as perception). They grant that Bayesian account of cognition can systematise the observational statements about the behaviour of cognisant organisms. They also grant that the Bayesian approach provides informative predictions about subjects' perceptual performance and functioning of their neural mechanisms. However, Colombo and Series deny that Bayesian inferences specify neuronal mechanistic or can be identified

[1] In my interpretation, they are referring to how actually models. The distinction between how possibly models of mechanisms and how actually models of them is important. The role of components must be identified by forming how actually models of phenomena (Craver 2007, 111 ff.). How possibly models are only heuristically useful. They could be used to specify the underpinning mechanisms loosely and without indicating that the components actually exist and contribute to the *explanandum* phenomena.

with them. On such grounds, Colombo and Series embrace an instrumentalist interpretation of the Bayesian theories of cognition (Colombo and Series 2012, 705). According to them, the Bayesian approach does not contribute to providing causal models of neural underpinning mechanisms of cognition. Therefore, despite their instrumental merits, Bayesian models cannot be identified with mechanisms that underpin the patterns of regularity and cannot provide plausible explanations of cognitive phenomena. Variations of the same theme have been developed by Colombo and Hartmann (Colombo and Hartmann 2015), who have argued that reliance on the mathematical relations drawn by Bayesian decision theory is not sufficient for unification in cognitive sciences. They acknowledge that contemporary neuroscience relies on the unifying power of Bayesian statistics to cover a wide range of phenomena from cue integration in perception, to sensorimotor control, causal learning and even social decision-making. They are also aware that the unifying power of Bayesian mechanisms consists in their ability to capture patterns of regularities through a few mathematical equations. Nonetheless, they claimed that the applicability of unifying mathematical models does not warrant a conclusion about the causal unification of phenomena.

This negative assessment of the unificatory powers the Bayesian theories of cognition could also apply to FEP and PEM. FEP and PEM are well-posed versions of the Bayesian approach to cognition. Accordingly, the same objections (to the unificatory claims of Bayesian cognitive science) could be extended to target the unificatory pretences of PEM too. Indeed, Colombo and Wright (Colombo and Wright 2016) have criticised the unificatory claims of the free energy formulation of PEM on such basis. This criticism draws its force from the inconsistency between the unificatory claims of FEP and the paradigm of mechanistic explanations (which are not committed to unificatory virtues). The critics submit that the "intended form of explanation afforded by PTB [aka FEP and PEM] is mechanistic rather than reductionistic" (Colombo and Wright 2016, 6). Then they argue that the fact that FEP enjoys a mathematical formulation that relates to other theories or subsumes them is not enough for attributing genuine (mechanistic) unifying powers to FEP. The same insight underlies Colombo and Wright's (2018) recent evaluation of the unificatory claims of FEP. Again, they pinpoint the inconsistency between FEP's unificatory power and the criterion of the plausibility of mechanistic explanations. According to them, "FEP is inconsistent with mechanistic approaches in the life sciences, which eschew laws and theories and take the explanatory power of scientific representations to be dependent on the degree of relevant biophysical detail they include" (Colombo and Wright 2018, 3). Then they deny FEP's unificatory claims.

What strikes me as odd is that Colombo and colleagues insist on evaluating the unificatory pretences of the Bayesian theories (and the implementation accounts such as FEP and PEM) from a mechanistic perspective, although they are well aware that these theories are inconsistent with the mechanistic paradigm. Colombo and Wright openly assert that "FEP is inconsistent with mechanism along two dimensions of representation: the dependence of explanatory force on describing mechanisms and the rejection of the idea that life science phenomena can be adequately explained through an axiomatic, physics-first approach" (Colombo and Wright 2018, 18). If the inconsistency is so explicit, why to dwell upon taking NMP as the right stance for evaluating FEP's claims, instead of taking a more charitable philosophical model. While I do not

pretend to know the reply to this last question, I assert that the main problem with this negative mechanistic evaluation of FEP's claims is that it takes the supremacy of the mechanistic explanations for granted, despite explicit awareness of the inconsistency between the paradigm of NMP and FEP. I do not intend to argue that the mechanistic approach cannot be applied to FEP, much less that NMP is a false philosophical theory. Rather, my point is that the supremacy of the mechanistic approach to explanation should not be taken for granted and that there are some alternatives which may provide a more charitable evaluation of the unificatory powers of FEP and PEM (and Bayesian cognitive science in general). In the next section, I flesh out such a charitable model and use it to evaluate FEP's unificatory pretences.

5 Ontic Structural Realism and Metaphysical Underdetermination

As I argued in the previous section, there was a mechanistic insight behind Colombo and colleagues' negative appraisal of the unificatory power of Bayesian cognitive science and FEP. The sort of unification that is at issue in Bayesian cognitive science is mainly based on mathematical relations. The applicability of Bayesian equations to different sorts of phenomena is not sufficient for accomplishing the goal of unification, in the sense that is at issue in NMP. Because unifying mathematical patterns are not identifiable in terms of causal mechanisms, they do not warrant a causal unification of the phenomena. It follows that the constraints that Bayesian models put on phenomena do not contribute to revealing mechanistic patterns of unification (although such constraints may have some heuristic value). This negative assessment of the unificatory power of FEP is justifiable but only from the perspective of a specific model of science. Taking an alternative perspective may (and indeed will) result in quite a different (i.e., a more charitable) evaluation of the unifying power of Bayesian cognitive science in general and FEP in particular. In other words, I argue that it is possible to vindicate unificatory claims of FEP[2], provided that we concede a structural realist approach to the evaluation. I do not justify structural realism (SR) in this paper, nor do I presume that it is the only correct philosophical model of science. However, I presume that the fact that the

[2] My vindication of the unificatory pretences of FEP (or the free energy formulation of PEM) presumes a representationalist, or moderate embodied construal but it does not presume that this is the exclusively correct construal of PEM and FEP. There are also radical embodied and enactivist construal of PEM. The embodied construal is presented in reaction to representationalism (which will be surveyed in the next section). The embodied approach denies that "our most elementary ways of engaging with the world" are representational (Hutto and Myin 2013, 13). The embodied cognition thesis recommends dispensing with the chasm between external features of the world and the internal symbolic representations of the cognitive agent (Varela et al. 1991). Such radical views inspire a radical embodied construal of PEM (Bruineberg and Rietveld 2014; Gallagher and Allen 2016). The embodied approach lays emphasis on the dynamical coupling of the organism with the environment and defines "agent" and its "environment" as a coupled entity. I suspect that an embodied construal of PEM is in line with explanatory pluralism. While this claim is worth discussing in a separate space, in this paper, I won't consider the bearing of the embodied construal of PEM on the discussion of its unifying power.

structural realist approach could rationally reconstruct and justify the unificationist claims of the advocates of FEP may tilt the balance in SR's favour.

SR is a moderate version of scientific realism. It is moderate in the sense that unlike full-blown scientific realism, it is not epistemically/ontically committed to individual objects but to structures. There are various forms of SR. Inspired by works of Henri Poincaré, John Worrall (Worrall 1989) presented an Epistemic version of SR (ESR) according to which our scientific theories contain *structural knowledge*. ESR successfully faces antirealist challenges that are based on theoretical changes in the history of science. There is also an ontic version of SR (OSR). According to OSR, the structure is all that there is (Ladyman 1998). OSR aims to defend a structural form of scientific realism in the face of metaphysical underdetermination caused by the diversity of theories of the modern physics that apply to the same field with equal empirical adequacy. Different formulations of theories of modern physics are committed to different kinds of (individual vs. non-individual) objects at the sub-particle level. The diversity of theoretical commitments in a field might wreak havoc with the thesis of realism because it might imply that theories do not provide a consistent picture of the unobservable parts of the world. The antirealist builds upon such cases to argue that the success of theories does not warrant their veracity. OSR can face this challenge. By arguing that there are common underlying structures (commonalities) that underpin theoretical diversities and unifies them, OSR defends a modified form of realism that is committed to structures, instead of individual objects.

It should be noted that the theoretical diversity that occasions metaphysical underdetermination in the ontology of physics somewhat resembles the theoretical diversity in the field of cognitive sciences. Theoretical diversity in the field of cognitive sciences could result in a state of explanatory pluralism. It has been argued that "Human beings, and in general the behavior and structure of complex biological entities, may not admit of one single theoretical paradigm for explanation" (Dale et al. 2009, 741). It follows that theoretical diversity of the field of cognitive science does not allow for unique specification (or identification) of cognitive mechanisms and processes. In this respect, explanatory pluralism (in cognitive science) and metaphysical underdetermination (in physics) are similar. In physics, too, theories of quantum statistics indicate that there are diverse arrangements of particles over states. For example, assuming that we have two particles and two states, one arrangement indicates that there is one particle in one state, whereas another arrangement leaves room for particle switching states (French 2014, 34–35). In the philosophy of physics, too, it would be quite hard to defend scientific realism in the face of diverse theoretical conceptions of the identity of objects. Ontic Structural Realists such as French and Ladyman endeavour to constrain the theoretical diversity (of the physics) and face the challenge of underdetermination (as a base for antirealism) by regimenting scientific theories into a meta-theoretical mathematical structure that plays a representational role. They argued that a common structure underlies diversities that jeopardise the thesis of realism, and such commonalities provide fresh unificationist grounds for defending (structural) scientific realism (French and Ladyman 2003). This is in line with Ladyman and Ross' statement of the metaphysical core of OSR, according to which metaphysics of SR consists in the unification of sciences (Ladyman and Ross 2007). As I say, the state of metaphysical underdetermination in physics bears

resemblance to cases of explanatory pluralism in cognitive sciences. Accordingly, it would be possible to face the challenge of pluralism in the field of cognitive sciences by a structural realist strategy (Beni 2016; Hasselman et al. 2010). I shall canvass some of such attempts in the next section.

In end this section with a quick comparison of agendas of OSR and NMP. In the previous section, I argued that the negative evaluation of the unificatory powers of PEM might have its roots in the mechanistic model that the critics have adopted. In this section, I showed that there is an alternative philosophical perspective which may provide a more charitable evaluation of the unificatory claims of PEM. NMP is accompanied by some realist assumptions about the nature of mechanisms that underpin the phenomena (see Craver 2014). NMP might be associated with a realist view on the nature of mechanisms that underpin explanations. SR, on the other hand, is not initially a theory of explanation (it is a realist theory), and in this respect, it is not only different to NMP but also the unificationist approach to explanation. This is because SR does not primarily aim to subsume the greatest number of explananda under the fewest number of argument patterns, but it aims to show how reliance on structural commonalities helps to go beyond underdetermination of ontology by theories. However, once we acknowledged the foundational status of the underpinning structures, we might as well rely on structures to develop explanatory paradigms. Decades ago, McMullin argued that properties and behaviours of some complex systems must be explained in virtue of their structures (McMullin 1978). More recently, some structure realist argued that well-known phenomena such as length contraction and time dilation (in relativistic physics) must be explained in virtue of the structural properties, namely the geometrical properties of Minkowski space-time (Dorato 2017; Dorato and Felline 2010). Thus, SR can lead to some structural accounts of explanation, and there are indeed points where mechanistic and structuralist approaches in the philosophy of science meet one another (Bechtel 2017; Felline 2015). Our enterprise in this paper could be regarded as another instance of confrontation of mechanistic and structural views. I will show that FEP (and PEM) provide a global explanation of life and cognition on the basis of unifying structure of the Bayesian meta-theoretical framework, rather than diverse biophysical mechanisms which could be modelled via varying approaches and methods.

6 Dealing with Diversities at the Meta-theoretical Level

OSR aims to overcome the state of metaphysical underdetermination. To do so, it relies on the unifying power of mathematical structures that play a representational role at a meta-theoretical level. The structures can be specified formally (e.g., in terms of set/model theory). This point explains why OSR's model may lead to a more favourable assessment of the unifying power of FEP, which enjoys an all-inclusive Bayesian formalism. In this section, I unpack my point with an eye to French's version of OSR, and in the next section, I will evaluate the unifying pretences of FEP from the perspective of structuralist models.

According to some ontic structural realists such as Steven French (and James Ladyman on several occasions), it is possible to rely on the unifying power of representational (mathematical) structures at a meta-theoretical level to overcome the state of metaphysical underdetermination caused by theoretical diversity in the field of physics. French and Ladyman argue that the representational structures could be characterised in terms of set/model theory (French and Ladyman 1999). This does not mean that structural realists are committed to the existence of set/model-theoretical structures as basic ontological units. Mathematical structures play only a representational role, and they are not ontologically constitutive (French 2014, 10). To flesh out their point, structural realists mark a distinction between mathematical structures that play a representational role and physical structures that are ontologically constitutive. Although physical structures are represented by unifying mathematical structures, they are not supposed to be ontologically reduced to the unifying mathematical structure. Unifications take place at a meta-level, i.e., the level of set/model theoretical structures. Mathematical representational structures are not identifiable with the physical structures that play a constitutive role in the ontology. However, structural relations form viable unifying patterns, and thereby a structuralist strategy could address the problem of the theoretical diversity and its diverse expressions, e.g., the problem of underdetermination or the problem of historical shifts.

I have to acknowledge that this move—i.e., conferring representational power upon the mathematical structures that are characterised at a meta-level—is not conceded by all structural realists. John Worrall's (1989) epistemic version of SR does not rely on meta-level set/model-theoretic structures. Instead, it deals with the issue of structural similarity at the level of the formalism of theories (e.g., differential calculus of theories of optics and electromagnetics). In the same vein, Margaret Morrison argues that there is no simple way to make this distinction between representational structures and theoretical (and physical) structures. She substantiates her point on the basis of Maxwell's equation, by arguing that the same mathematical notations that are invoked by the theory may also warrant unification (Morrison 2000). Similarly, there are versions of minimal scientific structuralism that presume that the shared structure of scientific theories—specified group-theoretically at the theoretical level—are sufficient for fleshing out the representational commitments of theories (Landry 2007). However, some notable versions of OSR—e.g., Steven French's version—make use of an advanced form of model theory[3] to specify representational commitments of theories at a meta-theoretical level. As French argues, theoretic structures cannot constrain the plurality of multiple representations and unify them at the level of scientific practice with enough clarity and precision (see French 2014, 105 ff. and Sect. 5.6). The Set/model-theoretic meta-theoretical framework of French's version of OSR enables it to constrain the theoretical diversity of scientific representation clearly.

To summarise, some notable versions of OSR presume that it is possible to regiment the representational structure of scientific theories at a meta-theoretical level that could be characterised mathematically, say, in terms of set/model theory (Bueno et al.

[3] It makes use of partial structures and partial isomorphisms, developed by French and colleagues) (Bueno et al. 2012; da Costa 2003).

2002; French and Ladyman 1999; Suppes 1967). This meta-theoretical formal regimentation enables OSR to emphasise the commonality between diverse theoretical implications of theories of modern physics and unify them. The meta-theoretical mathematical framework could unify diverse scientific representations without being ontologically constitutive.

Although, as my short survey of the origin of SR evinces, SR has been originated in the field of the philosophy of physics, there have been previous attempts at using a structural realist strategy for overcoming pluralism in the field of life sciences and psychology. I have previously advocates a structural realist strategy for overcoming pluralism caused by the theoretical diversity of neuroscientific accounts of the self (Beni 2016, 2018a). He uses a meta-theoretical framework (characterised information-theoretically) to unify diverse self-patterns that realise different aspects of selfhood, without implying that the selfhood is ontologically reducible to anything like informational structures. The meta-theoretical informational structure of the self is not ontologically constitutive, and yet it underpins the diversity of representations of aspects of the self. The structural realist strategy helps us to systematise vagaries of self-concepts and find a handle on a unified conception of the self. In life sciences, too, Steven French (French 2011, 2013) used a structural realist strategy to overcome the state of underdetermination that haunts the attempts at cutting the phylogenetic tree into right biological kinds. It has been argued that in biology, there are a number of different strategies for dealing with the question of how to specify biological kinds, and there are different replies to the general question of how to individuate biological systems which seem to be "massively integrated and interconnected" organisms (French 2011). Under the circumstances, French offered to deal with the problem of the *disunity* of approaches by invoking a structural realist strategy. He relied on high-level model-theoretical structures to unify diverse theoretical enterprises in the field of biology. French model-theoretic structuralism deals with biological pluralism by regimenting biological structures in terms of model-theoretic structures. He developed this approach to show how the model-theoretic framework underpins the trajectory of theoretical shifts from Mendel's laws of inheritance to theories of chromosome inheritance. A meta-theoretical model-theoretic framework could subsume diverse ways of individuating biological natural kinds and identifying units of selection (French 2011, 166, 2013). Inspired by such preceding attempts at extending SR to the field of special sciences, in the next section, I argue that OSR provides a good philosophical stance for (charitably) evaluating FEP's capacity for constraining the theoretical diversity of different phenomena (action, perception, and learning) and unifying various mechanisms whose functioning generates cognitive phenomena. FEP's unifying power does not need to be explicable in terms of causal mechanisms. From OSR point of view, FEP can provide global explanations of life and cognition in virtue of its comprehensive Bayesian framework, which subsumes varying approaches and hypothesis at a meta-theoretical formal level.

7 Evaluating the Unificatory Pretences of FEP

Being inspired by SR's success in the philosophy of physics (and recent extensions to the field of cognitive and life sciences), in this section I argue that it is possible to adopt a structural realist stance so as to support the unificatory pretences of FEP. This structural realist approach acknowledges the reliance of FEP's unificatory pretences on its comprehensive formal, Bayesian framework. This does not mean that we could rely on the Bayesian formalism to explain cognition in the same way that mechanists desire. FEP represents the basic structure of cognition at a meta-theoretical level that is characterised in terms of Bayesianism, and thereby, FEP explains some fundamental features of life and cognition globally, despite the fact that free energy "is not a directly measurable physical quantity" (Buckley et al. 2017, 57). The general insight of my structural realist approach is that unless we could elucidate the connection between diverse cognitive mechanisms, we would not be able to produce global and comprehensive explanations of cognition.

Pluralists such as Colombo and Wright assert that "understanding phenomena in the life sciences should allow for incompatible approaches with varying degrees of idealization" (Colombo and Wright 2018, 3). The structuralist approach allows us to subsume the incompatible approaches with different degrees of idealisation and various mechanisms that feature in them and incorporate them into global explanations of cognitive phenomena. Attributing unifying powers to FEP is compatible with accepting the existence of various biophysical mechanisms (and varying approaches to modelling them). Various theories of implementations could explain some specific features of cognition on the basis of neurobiological and biophysical mechanisms but the scope of such explanations is limited. For example, it is possible to explain the brain's experience of joy (or lack thereof) on the basis of neurobiological mechanisms of dopaminergic catecholaminergic neurotransmitter release that are implemented in the ventral pallidum (VP) and rostromedial shell of the nucleus accumbens (NAcc) (Colombo and Wright 2016, 7; Pecina and Berridge 2005). The feeling of joy or the anhedonia can be explicated on the basis of (positive or negative) functional relations between component parts and operations of the nervous system (e.g., VP, NAcc, and dopaminergic system), but the produced explanations would be local. Such limited explanations do not produce insights into the global nature of cognition in general. Similarly, theories of incentive salience aim to explain how the operation of the mesocorticolimbic dopaminergic system causes the cognitive system's function, or its ability to represent some external stimuli as more salient, attractive or desirable (Berridge 2012). However, the explanatory scope of this specific theory, too, is limited. By unifying such local theories and approaches at a meta-theoretical level, FEP produces comprehensive explanations of global features of cognition and life, e.g., about how the organism interacts with its environment. On this subject, please note that explanans and the explanandum that feature in FEP-based account of cognition are the same regardless of whether we adopt a structural realist or a mechanistic stance. The explanans is the free energy principle, i.e., a unifying fundamental law regimented in terms of a Bayesian framework with a unifying power. Explananda are various facts such as the organism's successful interaction with its environment, maximisation of

survival, successful cognition, perception, and action, etc. FEP explains all of these facts on the basis of a fundamental law, i.e., free energy principle, and in this sense, it comes with great unificatory pretences. In this context, while NMP, which only recognises the plausibility of mechanistic explanations, submits a negative evaluation of FEP's unificatory power, SR provides a viable stance for defending a sympathetic understanding of how FEP explains various types of phenomena by subsuming them under the rubric of the free energy principle. So, the difference is in the respective evaluations that NMP and SR provide of FEP's unificatory pretences.

Be that as may, FEP provides a powerful meta-theoretical framework for subsuming diverse strategies, experimental studies, and hypothetical stipulations concerning, say, cue integration, perception, categorization, emotion, selfhood, etc. The structural realist perspective acknowledges the explanatory power of unifying structures, and it submits that we can account for the versatility of cognition and its relation to life in virtue of Bayesian inferences that the brain forms to model the world.

From the structural realist point of view, FEP does not provide a mechanistic sketch, i.e., an incomplete representation of a mechanism that should be filled with details (Colombo and Wright 2018; Piccinini and Craver 2011). From NMP's point of view, Sketches are rudimentary, incomplete attempts at describing a mechanism. Mechanistic sketches can obtain the status of full-blown mechanistic explanations if we add the omitted mechanical details about component parts and operations. NMP presumes that mechanistic sketches as such do not provide viable and complete explanations. My structural realist approach, on the other hand, submits that it is possible to account for some fundamental features of cognition on the basis of the unifying power of formal structures. I will unpack this remark immediately.

I agree with critics of the unifying power of FEP such as (Colombo and Wright 2016, 2018) about the fact that FEP's unifying power is based on the capacity of mathematical structures that cover diverse theories of perception, cognition, life, action and social decision making. However, while Colombo and colleagues (who were taking a mechanistic stance) presumed that the Bayesian framework of theories of cognition could not exert genuine unifying-explanatory power, my structural realist stance allows for admiring the unificatory power of FEP, which emerges in virtue of its Bayesian meta-theoretical framework. The Bayesian framework includes unifying structures that contribute to the global explanation of cognition in a way that remains beyond the scope of the mechanistic approach. It is in virtue of the organisms' capacity to minimise their amount of surprise that FEP can show how *all systems* with homoeostatic (autopoietic) properties—i.e., systems capable of self-organisation and reproduction—stay in the state of equilibrium with their environment and resist the tendency to disorder by putting an upper bound on their internal entropy. The organisms' capacity to minimise their surprise could be explained not on the basis of the operation of diverse mechanisms that constitute each organism. But we cannot build our global account of features of cognition on the basis of diversified mechanisms. We could explain how the organism could resist the dispersing effect of the environment and get a cognitive handle on the structure of the world in virtue of the variational Bayesian framework that the organism employs to model the states of the environment and optimise its models of those states.

FEP delineates the most fundamental principle of the organism-environment relationship on the basis of the mathematical formulation of notions of free energy, surprise, entropy, and active inference. PEM is a special case of the FEP in the field of cognition. According to PEM, in order to minimise its prediction error, the organism must be able to model the causal structure of the environment to itself efficiently. The error reducing capacity of the brain could be basically explicated on the basis of the variational Bayesian models that the brain invokes to predict the states of the world and optimise its models of those states. A single formal framework, i.e., variational Bayesianism, subsumes the structures of diverse mechanisms that contribute to the formation of the organism's representation of the causal structures in the environment. There are of course varying mechanisms that contribute to the formation of the brain's representation at the physical and computational level. However, my point is that the most global explanation of the organism's capacity to maximise its survival in the changing environment (and the brain's capacity to model the causal structure of the environment) can be offered in terms of the variation Bayesian formalism that models the foundational principles of organism's survival and cognition.

It is true that, at the theoretical level, various approaches and hypothesis could be used to model (biophysical and computational) mechanisms of the brain's activity. But the explanations that this approach accommodates are local and limited. Variational Bayesianism regiments the models of cognition, action, and representation of simple organisms, single brains and large-scale processes that take place across large spatiotemporal structures—e.g., evolution[4]— at a meta-theoretical level. The meta-theoretical framework glosses over mechanical details and thereby provides a global explanation of the organism's relationship with the environment (and the brains' cognitive powers). The critics of the unificatory powers of PEM (and FEP) submit that the theory lacks explanatory power because it disregards the "biophysical reality of the nervous system" (Colombo and Wright 2018, 2). My point is that, because FEP glosses over mechanical details, it can globally explain fundamental facts about how the organism maximises its survival and (given PEM derivation) how the brain represents the structure of the world. Please note that even the critics of FEP's unifying power are clear that FEP endeavours to explain various phenomena such as "Hebb's rule and spike-timing dependent plasticity, the multiplicity and hierarchical organization of cortical layers" not mechanistically, but "axiomatically, as logical deductions from sets of axioms and formulae" (Colombo and Wright 2018, 18). This means that even critics such as Colombo and Wright implicitly consent that FEP can accommodate structural explanations. However, the critics' presupposition about the supremacy of mechanistic explanations deters from accepting the legitimacy of structural explanations. SR legitimises the structural paradigm of explanation. It holds that the organism stays in homeostatic states or states of equilibrium with the fluctuating environmental conditions by putting an upper bound on the entropy (or total surprisal) of the system and in virtue of Bayesian inferences that it implements. Cognisant organisms represent the

[4] According to this reading, natural selection can be construed as a Bayesian model selection process based upon the adaptive fitness—that is scored by the surprise accumulated by a phenotype (see Allen and Friston 2016).

causal structure of the world by implementing the same equations. It follows that the meta-theoretical Bayesian framework accommodates global explanation of adaptation, cognition, and survival. This seems to be in complete agreement with French's approach to OSR according to which the regimentation of scientific representations at the meta-theoretical level enables us to overcome the state of diversities and indeterminacies in the fields of physics and life sciences.

8 Concluding Remarks

Two dominant approaches of the philosophy of science, NMP and SR, rephrase scientific theories in terms of two different kinds of (mechanistic and structural) models. In this paper, I emphasised this point to show that the question of unificatory pretences of FEP should not receive a straightforward reply. Pluralists and unificationists endeavoured to either criticise or praise the unificatory power of FEP monolithically and as a factive statement. In this paper, I aimed to show that the issue of FEP's unificatory power is more subtle than what may appear at first glance. I also argued that, while the mechanistic approach motivated some negative evaluation of the unificatory powers of FEP, a structural realist perspective may bear more favourably on our evaluation of these unificatory powers.

In this paper, I showed that the critics of the unifying power of FEP, such as Colombo and colleagues, usually use mechanistic models to evaluate FEP's unificatory claims. It has been contended that because the Bayesian framework of FEP cannot be identified with the underpinning mechanisms, its unificatory power is not genuine (from a mechanistic point of view). I argued that it is possible to use structural models when evaluating the unificatory claims of FEP. This perspective recognises the role of common underlying structures in unifying diverse theoretical and experimental aspects of theories. The unifying structures can be specified not at the level of scientific theories or scientific practice, but at a high level of abstraction, i.e., at the level of meta-theoretical frameworks. In this paper, I suggested that the meta-theoretical framework of FEP could be characterised in terms of variational Bayesianism. I argued that it is in virtue of this meta-theoretical framework that we can accommodate global explanations of the organism's capacity to minimise the discrepancy between its models and the environment.

References

Alderson-Day B et al (2016) Auditory hallucinations and the brain's resting-state networks: findings and methodological observations. Schizophr Bull. http://www.ncbi.nlm.nih.gov/pubmed/27280452. Accessed 21 July 2016

Allen M, Friston KJ (2016) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese 1–24. http://link.springer.com/10.1007/s11229-016-1288-5. Accessed 14 Dec 2017

Bechtel W (2017) Analysing network models to make discoveries about biological mechanisms. Br J Philos Sci. https://academic.oup.com/bjps/advance-article/doi/10.1093/bjps/axx051/4100193. Accessed 11 June 2018

Bechtel W, Abrahamsen A (2005) Explanation: a mechanist alternative. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 36(2):421–441

Beni MD (2016) Structural realist account of the self. Synthese 193(12):3727–3740. http://link.springer.com/10.1007/s11229-016-1098-9. Accessed 3 Dec 2016

Beni MD (2018a) An outline of a unified theory of the relational self: grounding the self in the manifold of interpersonal relations. Phenomenol Cogn Sci 1–19. http://link.springer.com/10.1007/s11097-018-9587-6. Accessed 28 July 2018

Berridge KC (2012) From prediction error to incentive salience: mesolimbic computation of reward motivation. Eur J Neurosci 35(7):1124–1143. http://www.ncbi.nlm.nih.gov/pubmed/22487042. 11 Dec 2016

Bruineberg J, Rietveld E (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. Front Hum Neurosci 8: 599. http://journal.frontiersin.org/article/10.3389/fnhum.2014.00599/abstract. 21 June 2017

Buckley CL, Kim CS, McGregor S, Seth AK (2017) The free energy principle for action and perception: a mathematical review. J Math Psychol 81: 55–79. https://www.sciencedirect.com/science/article/pii/S0022249617300962. Accessed 17 Sept 2018

Bueno O, French S, Ladyman J (2012) Models and structures: phenomenological and partial. Stud Hist Philos Sci Part B: Stud Hist Philos Mod Phys 43(1):43–46. http://www.sciencedirect.com/science/article/pii/S1355219811000712. Accessed 25 Oct 2017

Bueno O, French S, Ladyman J (2002) On representing the relationship between the mathematical and the empirical. Philos Sci 69:497–518

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36(03):181–204. http://www.journals.cambridge.org/abstract_S0140525X12000477. Accessed 8 June 2016

Colombo M, Hartmann S (2015) Bayesian cognitive science, unification, and explanation. Br J Philos Sci 68(2):axv036. http://bjps.oxfordjournals.org/lookup/doi/10.1093/bjps/axv036. Accessed 17 Oct 2018

Colombo M, Series P (2012) Bayes in the brain–on bayesian modelling in neuroscience. Br J Philos Sci 63(3):697–723. http://bjps.oxfordjournals.org/cgi/doi/10.1093/bjps/axr043. Accessed 25 July 2016

Colombo M, Wright C (2016) Explanatory pluralism: an unrewarding prediction error for free energy theorists. Brain Cogn. http://www.ncbi.nlm.nih.gov/pubmed/26905647. Accessed 10 Dec 2016

Colombo M, Wright C (2018) First principles in the life sciences: the free-energy principle, organicism, and mechanism. Synthese 1–26. http://link.springer.com/10.1007/s11229-018-01932-w. Accessed 25 Dec 2018

da Costa, NCA, French S (2003) Science and partial truth. Oxford University Press, Oxford. http://www.oxfordscholarship.com/view/10.1093/019515651X.001.0001/acprof-9780195156515. Accessed 29 May 2016

Craver CF (2007) Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Clarendon Press, Oxford

Craver CF (2014) The ontic account of scientific explanation. In: Explanation in the special sciences. Springer, Dordrecht, pp 27–52. http://link.springer.com/10.1007/978-94-007-7563-3_2. Accessed 26 Nov 2016

Dale R, Dietrich E, Chemero A (2009) Explanatory pluralism in cognitive science. Cogn Sci 33 (5):739–742

Dorato M (2017) Dynamical versus structural explanations in scientific revolutions. Synthese 194(7): 2307–2327. http://link.springer.com/10.1007/s11229-014-0546-7. Accessed 2 Sept 2017

Dorato M, Felline L (2010) Structural explanations in minkowski spacetime: which account of models? In: Space, Time, and Spacetime. Springer, Heidelberg, pp 193–207. http://link.springer.com/10.1007/978-3-642-13538-5_9. Accessed 21 Nov 2016

Felline L (2015) Mechanisms meet structural explanation. Synthese 1–16. http://link.springer.com/10.1007/s11229-015-0746-9. Accessed 21 Nov 2016

French S (2011) Shifting to structures in physics and biology: a prophylactic for promiscuous realism. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 42(2):164–173. https://www.sciencedirect.com/science/article/pii/S1369848610001160. Accessed 8 Sept 2018

French S (2013) Eschewing entities: outlining a biology based form of structural realism. In: EPSA11 perspectives and foundational problems in philosophy of science. Springer, Cham, pp 371–381. http://link.springer.com/10.1007/978-3-319-01306-0_30. Accessed 10 July 2018

French S (2014) The structure of the world: metaphysics and representation. Oxford University Press, Oxford. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199684847.001.0001/acprof-9780199684847. Accessed 12 Oct 2017

French S, Ladyman J (1999) Reinflating the semantic approach. Int Stud Philos Sci 13(2):103–121. http://www.tandfonline.com/doi/abs/10.1080/02698599908573612. Accessed 27 Dec 2016

French S, Ladyman J (2003) Remodelling structural realism: quantum physics and the metaphysics of structure. Synthese 136(1): 31–56

Friston KJ (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11 (2):127–138. http://www.ncbi.nlm.nih.gov/pubmed/20068583. Accessed 13 July 2016

Friston KJ, Stephan KE (2007) Free-energy and the brain. Synthese 159(3):417–458. http://www.ncbi.nlm.nih.gov/pubmed/19325932. Accessed 8 Jan 2017

Friston KJ, Thornton C, Clark A (2012) Free-energy minimization and the dark-room problem. Front Psychol 3:130. http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00130/abstract. Accessed 24 July 2016

Gallagher S, Allen M (2016) Active inference, enactivism and the hermeneutics of social cognition. Synthese 1–22. http://link.springer.com/10.1007/s11229-016-1269-8. Accessed 14 Dec 2017

Glennan S (2002) Rethinking mechanistic explanation. Philos Sci 69(S3):S342–S353. http://www.journals.uchicago.edu/doi/10.1086/341857. Accessed 26 Nov 2016

Glennan S (2017) The New Mechanical Philosophy. Oxford University Press, Oxford. http://www.oxfordscholarship.com/view/10.1093/oso/9780198779711.001.0001/oso-9780198779711. Accessed 18 July 2018

Godfrey-Smith P (2009) Models and fictions in science. Philos Stud 143(1):101–116. http://link.springer.com/10.1007/s11098-008-9313-2. Accessed 30 Oct 2016

Hasselman F, Seevinck MP, Cox RFA (2010) Caught in the undertow: there is structure beneath the ontic stream. SSRN Electron J. http://www.ssrn.com/abstract=2553223 Accessed 23 May 2016

Hohwy J (2013) The predictive mind. Oxford University Press, Oxford. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199682737.001.0001/acprof-9780199682737. Accessed 30 May 2016

Hohwy J (2014) The self-evidencing brain. Noûs 50(2):259–285. http://doi.wiley.com/10.1111/nous.12062. Accessed 30 May 2016

Horga G, Schatz KC, Abi-Dargham A, Peterson BS (2014) Deficits in predictive coding underlie hallucinations in schizophrenia. J Neurosci Off J Soc Neurosci 34(24):8072–8082. http://www.ncbi.nlm.nih.gov/pubmed/24920613. Accessed 21 July 2016

Hutto DD, Myin E (2013) Radicalizing Enactivism Basic Minds without Content. MIT Press, Cambridge

Kaplan DM (2011) Explanation and description in computational neuroscience. Synthese 183 (3):339–373. http://link.springer.com/10.1007/s11229-011-9970-0. Accessed 29 Nov 2017

Kitcher P (1989) Explanatory unification and the causal structure of the world. In: Kitcher P, Salmon W (eds) Scientific Explanation. University of Minnesota Press, Minneapolis. https://conservancy.umn.edu/handle/11299/185687. Accessed 22 Dec 2018

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27(12):712–719

Ladyman J (1998) What is structural realism? Stud Hist Philos Sci Part A 29(3):409–424

Ladyman J, Ross D (2007) Every thing must go. Oxford University Press, Oxford

Landry EM (2007) Shared structure need not be shared set-structure. Synthese 158(1):1–17. http://link.springer.com/10.1007/s11229-006-9047-7. Accessed 25 Jan 2018

McMullin E (1978) Structural explanation. Am Philos Q 15:139–147. https://www.jstor.org/stable/20009706. Accessed 8 Dec 2018

Miłkowski M (2016) Unification strategies in cognitive science. Stud Logic Grammar Rhetoric 48(61):13–33. https://philpapers.org/archive/MIKUSI.pdf. Accessed 1 Aug 2017

Mitchell SD (2003) Biological complexity and integrative pluralism. Cambridge University Press, Cambridge

Morrison M (2000) Unifying scientific theories. Cambridge University Press, Cambridge. http://ebooks.cambridge.org/ref/id/CBO9780511527333. Accessed 17 Oct 2018

Pecina S, Berridge KC (2005) Hedonic hot spot in nucleus accumbens shell: where do μ-opioids cause increased hedonic impact of sweetness? J Neurosci 25(50):11777–11786. http://www.ncbi.nlm.nih.gov/pubmed/16354936. Accessed 11 Dec 2016

Piccinini G, Craver C (2011) Integrating psychology and neuroscience: functional analyses as mechanism sketches. Synthese 183(3):283–311. http://link.springer.com/10.1007/s11229-011-9898-4. Accessed 11 July 2017

Ramstead MJD, Badcock PB, Friston KJ (2017) Answering schrödinger's question: a free-energy formulation. Phys Life Rev. https://www.sciencedirect.com/science/article/pii/S1571064517301409. Accessed 2 Jan 2018

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87. http://www.ncbi.nlm.nih.gov/pubmed/10195184. Accessed 17 June 2016

Sanborn AN, Chater N (2016) Bayesian brains without probabilities. Trends Cogn Sci 20 (12):883–893. https://www.sciencedirect.com/science/article/pii/S1364661316301565. Accessed 22 Dec 2018

Seth AK (2014) A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. Cogn Neurosci 5(2):97–118. http://dx.doi.org/10.1080/17588928.2013.877880. Accessed 21 July 2016

Seth AK (2015) The cybernetic bayesian brain. Open-mind.net. http://www.open-mind.net/papers/@@chapters?nr=35. Accessed 3 Aug 2016

Suppes P (1967) What is a scientific theory? In: Morgenbesser S (ed) Philosophy of Science Today. Basic Books, New York, pp 55–67. https://www.google.com/_/chrome/newtab?espv=2&ie=UTF-8

Tabery J (2014) Beyond versus: the struggle to understand the interaction of nature and nurture

Tabery J, Preda A, Longino H (2014) Pluralism, social action and the causal space of human behavior. Metascience 23(3):443–459. http://link.springer.com/10.1007/s11016-014-9903-x. Accessed 26 Nov 2016

Varela FJ, Thompson E, Rosch E (1991). The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge

Weisberg M (2013) Simulation and Similarity Using Models to Understand the World. Oxford University Press, Oxford

Worrall J (1989) Structural realism: the best of both worlds? Dialectica 43(1–2):99–124. http://doi.wiley.com/10.1111/j.1746-8361.1989.tb00933.x



How Philosophical Reasoning and Neuroscientific Modeling Come Together

Gabriele Ferretti[1(✉)] and Marco Viola[2]

[1] Department of Philosophy, University of Florence, Florence, Italy
fairg@live.it
[2] Department of Philosophy and Education, University of Turin, Turin, Italy

Abstract. Is there any fruitful interplay between philosophy and neuroscience? In this paper, we provide four case studies showcasing that: (i) Philosophical questions can be tackled by recruiting neuroscientific evidence; (ii) the epistemological reflections of philosophers contribute to tackle some foundational issues of (cognitive) neuroscience. (i) will be supported by the analysis of the literature on picture perception and Molyneux's question; (ii) will be supported by the analysis of the literature on forward and reverse inferences. We conclude by providing some philosophical reflections on the interpretation of these cases.

1 Introduction

Several scholars acknowledge that there can be a rich and fruitful interplay between the empirical portion of cognitive science and the "part of cognitive science that is philosophy"[1] (e.g. Bechtel 2009; Brook 2009; Dennett 2009). However, while traditionally the empirical side of this debate was mainly driven by *cognitive psychology*, nowadays a pivotal role is played by *cognitive neuroscience* (henceforth, simply *neuroscience*). Only in recent times, philosophers showed that they recognize more and more the relevance of using empirical results from *neuroscience* to investigate *philosophical problems*. In the meantime, *philosophical speculation* has been put at work in investigating the *epistemological foundations* of the *neuroscientific practice*. In this respect, it seems that neuroscientific modeling and philosophical reasoning can massively interact. The former can offer empirical irons in the fire of the philosopher that embraces a naturalistic perspective when tackling genuine philosophical conundrums. The latter can provide conceptual tools that can regiment, from a logical and inferential point of view, the models proposed by neuroscientists as explanations of several neural phenomena.

In this paper, we want to move some steps toward an account of how *philosophical reasoning* and *neuroscientific modeling really interlock*. However, rather than beginning with *apriori* discussions, we will guide the reader through a brief tour, or round-trip, of some recent debates we are familiar with in our work of research, and that we think represent genuine cases of fruitful interactions.

[1] The expression is by Dennett (2009).

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 173–190, 2019.
https://doi.org/10.1007/978-3-030-32722-4_11

The first two cases show how experimental evidence drawn from neuroscientific models can help philosophical reasoning aimed at solving theoretical puzzles: this is the part of the trip that goes from neuroscience to philosophy.

The latter two cases suggest what is the contribution of philosophical (epistemological) reasoning to foundational problems related to neuroscientific modeling: this is the part of the trip that goes from philosophy to neuroscience.

These case studies hopefully show that (and how) a fruitful interlock between *neuroscientific modeling* and *philosophical reasoning* might obtain: philosophical questions can be purposefully reframed and refined thanks to empirical data, whereas neuroscientific issues cannot be solved by mere empirical data, because they set the rules for interpreting these data.

2 From Neuroscience to Philosophy

In this section, we propose two case studies the discussion of which will serve as illustration of how philosophers are effectively recruiting more and more experimental results from neuroscience in order to carry out their philosophical investigations. Not only does this methodology allow philosophers to offer more appropriate, empirically informed answers to their philosophical questions. Furthermore, this practice has the advantage of flagging them the best way to posit such questions in a way that is scientifically meaningful in the light of what neuroscience is teaching us about the brain, and philosophically interesting for those who embrace a naturalistic stance. We could sum up the stance behind this philosophical inclination to experimental results by following the wonderful words by Ned Block:

> (…) it would be a mistake to think that those who know nothing of the science of the mind can just stick to the relatively a priori parts of philosophy of mind, since one needs to understand the empirical facts to even know where there is room for relatively a priori philosophy (Block 2014: 570–571).

The first case study is about picture perception, whereas the second one concerns the famous Molyneux's question.

2.1 Picture Perception

What is the difference between perceiving a depicted apple and perceiving it in the flesh? Since the old debate about the nature of picture perception between Wollheim (1998) and Gombrich (1960), several philosophers have investigated the difference between these two perceptual states. Several scholars have suggested that an empirically-informed philosophical investigation would be very beneficial when it comes to the understanding of what is the answer to this question.

Following a naturalistic stance, it is possible to ask, in the light of what we know about human neurophysiology, what happens to our visual system in these two different situations. Such a question is not trivial, especially if we consider the presence of special cases of illusory picture perception. Notably, in the case of *trompe l'oeils*, we are in front of a depicted object which nonetheless looks, even if only momentarily,

like a real object. The challenge then becomes that of understanding how our visual brain can generate pictorial experience on the basis of specific visual stimuli, especially with respect to the case of strange, *illusory pictorial beasts* of perception that *trompe l'oeils* happen to be.

At this point, it may be clear to the reader that a good theory of picture perception should be inscribed within a theory of perception of presence of real objects. Indeed, understanding the difference between these perceptual states amounts to understanding how we visually represent some of the pieces of the visual world as real and present, while some others as merely pictorial.

Mohan Matthen has been the first to suggest, in his book (2005), that taking a look at our best neuroscientific theory of vision could have been useful for our understanding of the peculiarity of picture perception. With this spirit, he has proposed to take look at the famous *Two Visual Systems Model* proposed by Milner and Goodale (1995/2006), in order to say something informative about the behavior of our visual system when we are in front of pictorial objects.

According to the first formulation of this theory, visual consciousness and visuomotor control turn out to be subserved from visual areas that are different on an anatomo-functional level. In particular, this theory suggests that in humans and other mammals a separation obtains between a ventral stream involved in conscious visual recognition of the surrounding environment, and a dorsal stream for the visual guidance of action, whose processing is not accessible in our visual consciousness. The possibility of dissociation between these two pathways is attested by lesion studies related to the visual cortex. Lesions to the dorsal pathway (i.e., the occipito-parietal route that goes from the primary visual cortex, through the posterior parietal cortex, towards the premotor and motor areas) compromise the possibility of using visual information for the guidance of action (leading to what is known in the literature as *optic ataxia*), while leaving intact the visual processing responsible for conscious visual recognition. Lesions to the ventral pathway (i.e. occipito-temporal route that goes from the primary visual cortex, to the inferotemporal cortex) compromise the possibility of conscious recognition (leading to what is known in the literature as *visual agnosia*), while leaving intact the visual processing responsible for the guidance of action (Milner and Goodale 1995/2006). Such a dissociation is also suggested by behavioral studies which show that, in healthy individuals, some visual illusions can 'deceive' our conscious visual recognition, without, however, any sign of deception being exhibited by vision for action (Milner and Goodale 1995/2006).

The challenge of proposing an account of picture perception that is empirically informed by the *Two Visual Systems Model*, advocated by Matthen (2005), has been effectively met by Bence Nanay (2011), who has effectively built such a theory. According to Nanay, being the ventral pathway responsible for visual recognition, it has to be involved in the representation of the depicted object, as well as, in certain cases, also in that of the surface (when, for example, we enter an aesthetic appreciation of a pictorial piece of art and we are able to visually spot the majesty with which a mark on a surface is codified across a material surface, so that it can be visually encoded on it). However, since we cannot act on depicted objects, the dorsal pathway, which is responsible for the construction of our visuomotor responses, cannot respond to depicted objects, but only to real objects, which we can perceive as suitable for reliable interaction. This seems to be the

difference with the perception of *trompe l'oeils*, in which the dorsal response tracks the *presence for action* (Nanay 2015).

Following Ferretti (2016a, 2016c), however, such a theory could be extended taking into account empirical evidence showing that also the dorsal pathway responds to depicted objects, with the proviso that they are presented as apparently located in the action space of the observer. This, however, amounts to say that dorsal visuomotor processing cannot be responsible for discriminating between usual picture perception, *trompe l'oeil* perception and the perception of real objects. Thus, by offering a philosophical analysis of the most recent empirical evidence, Ferretti (2016a, 2016c, 2018, forthcoming) has proposed a theory capable of taking into account the fact that both the visual pathways respond to both the depicted and normal objects and that, at the same time, can explain what happens when we perceive *trompe l'oeils*. The story is the following.

The dorsal pathway lacks the computational resources allowing distinguishing between real and depicted objects. Such a visual task is subserved by the computations related to the visual recognition of the object, which are courtesy of ventral processing. Indeed, some dorsal neural populations, responsible for the computation of motor acts, are activated also when we perceive a depicted object whose geometrical properties can be translated in action properties[2]. Indeed, our motor representations can be attuned to depicted objects (Ferretti 2016b): when we are in front of a real object, there are dorsal computations that trigger the representations of the motor act that would be congruent with the object if this were a real one – even though dorsal vision *does not know that it is not real* (pun intended). This activation, however, is obtained only if the object is apparently presented within the peripersonal space of the observer – this is also because, at the cortico-cortical level, the activity of the AIP-F5 circuit which is responsible for the computation of motor acts, works in interplay with the VIP-F4 circuit, whose task is to compute the peripersonal space (Ferretti 2016b, 2016d). If so, why don't we act upon the depicted object? Well, this is because the plethora of computational interactions between ventral and dorsal processing, concerning visual recognition and action guidance, allow us to distinguish between a real and a depicted object. The ventral pathway can distinguish between different stimuli and can establish whether action planning can be activated or not, triggering, subsequently, dorsal computations for action. For this reason, it is (almost) impossible to have the intention to act on an object that the ventral pathway has recognized as depicted, even if dorsal responses are activated (Ferretti 2018).

Here is another important question. If ventral processing represents whether the object is real or not (i.e. it is depicted), why does dorsal processing give raise to the computations of the motor acts related to the object independently of whether it is real or depicted? This is for a simple reason. Dorsal processing has what is called a magno-cellular advantage. Dorsal responses, not by chance automatic, are extremely faster than those of ventral processing, which are linked to the parvo-cellular pathways (Milner and Goodale 1995/2006). Thus, when we look at an object, dorsal responses arrive before we can perform conscious visual recognition on the object. However, the

[2] For example, those pertaining to the parieto-premotor circuit AIP-F5, within the ventro-dorsal pathway.

result of dorsal computations, related to action processing, are available only after that ventral computations, related to conscious visual recognition, attested whether the object is real or depicted. Therefore, if we decide not to act, or if we recognize that we *cannot* act on a specific given object, dorsal responses remain stored in the motor memory and then decay (for a review see, Ferretti 2018).

All this seems to suggest that, with *trompe l'oeils*, it is the ventral pathway that is 'deceived' (pun intended), not being able to recognize the object as a pictorial one[3]. The difference between a normal pictorial experience and an illusory one depends on what it has been just explained and not, as previously supposed, in a different response concerning computational processing of the dorsal pathway, which is, as suggested, always active (for a review see, Ferretti 2018, forthcoming).

Furthermore, numerous experimental results seem to suggest that the functional dissociation between the streams, and their related activities, are not very deep: the functional activity of the dorsal pathways is always 'supported' by that of the ventral pathway and vice versa (Chinellato and Del Pobil 2016; Zipoli Caiani and Ferretti 2017; Ferretti 2018, Zipoli Caiani 2013).

In this respect, an empirical informed theory of pictorial perception should take into account such an evidence. In this respect, it has been recently suggested (Ferretti 2016c, 2018) how the functional interaction between the two streams is of an extreme importance, as ventral computations are never really '*purely* ventral', as well as dorsal computations are never really '*purely* dorsal': they always influence each other, for different tasks, in different contexts. Indeed, it is the functional interaction between the two streams that prevents us from acting upon images. This is always due to the computational processes described above. However, the reader should note that such computational processes are not the result of the activity of a single pathway; rather, of the computational communion between the two pathways. It is, thus, the collaboration between the two visual streams that gives us an accurate perception of both the real objects present for action and of pictorial objects, as well as the possibility of distinguishing between them. In this respect, it has been shown how this same piece of evidence allows us to build up a theory of perceptual presence capable of explaining how we visually represent some fragments of the external word as being real, while others not (Ferretti 2016c, 2017a, 2017b, 2017c, 2018, 2019b, forthcoming).

Finally, such a theory can explain a strange fact occurring in the literature on pictorial perception. On the one hand, the philosophical analysis seems to clearly show how both depicted objects and real objects can solicit our visual system in a very similar way. On the other hand, however, it is clear how these two perceptual states are different. In this respect, our visual system has evolved by encoding real objects, as pictures are artifacts arrived late in our evolution. Nonetheless, most of the neuroscientific studies involved in the investigation of how our visual system reconstruct the external world makes use of pictures of objects in the experimental settings built in the laboratories. On the one hand, we know that real and depicted objects are not very similar for perception. On the other, the latter are usually employed in order to

[3] There are here computational explanations for this fact, based on the functional activity of the streams, that cannot be analyzed in this venue. (see Ferretti 2016c, 2017c, 2018, 2019b)

understand how our visual brain works. It has been suggested (Ferretti 2017c) that a satisfying theory a picture perception should also have the power of explaining how we can maintain both the philosophical stance on picture perception, as well as the experimental practice described above, without incurring in methodological problems, both on the theoretical and on the experimental ground. This is possible only by offering a notion of pictorial perception that succeeds in taking into account how our visual pathways are at work when we perceive a real object and when we perceive a depicted object (Ferretti 2017c).

The aim of this section was to show that philosophers have effectively taken into account the results from vision neuroscience in order to investigate the nature of pictorial perception. In this venue it is not possible, for reasons of space, as well as for coherence of content, to offer a technical explanation of *how* they effectively used these experimental results. Namely, to suggest a theory that allows us to explain how the two visual streams, as well as the complex interactions between them, can lead us to visually represent, in an appropriate manner, both the picture's surface and the depicted object simultaneously. A theory that would make in tune our best philosophical story about picture perception, with our the best explanation of how our visual system works (for the most recent theory see Ferretti 2018).

2.2 Molyneux's Question

Here is Molyneux's question. Imagine a subject born blind, who learnt to discriminate specific shapes by using touch. Can she/he immediately recognize, in the case in which her/his vision is suddenly restored, the same shapes placed before her/his eyes by using vision?

The real point lurking behind this thought concerns whether the special link between vision and touch is the result of a perceptual learning given in our everyday experience of the external world, or, differently, such a link is already given in our sensory *repertoire*, without the need of any ontogenetic evolution (Degenaar and Lokhorst 2014; Schwenkler 2013).

Molyneux's question can be regarded as a genuine philosophical question on what are defined, in the philosophical literature, the contents of perception (Siegel 2010). However, several philosophers have suggested that, in order to properly understand how an answer to such a question can be cashed out, experimental evidence is crucial (Schwenkler 2013; Gallagher 2005; Jacomuzzi et al. 2003; Glenney 2013; Ferretti and Glenney, Under Contract).

Now, in order to build an appropriate experimental setting, we must have at our disposal congenitally blind subjects to whom vision will be subsequently restored. It is in this framework that a philosophical question, which investigates the horizon of vision and its nature, opens for us an empirical chink that can help us to understand whether it is possible to restore the visual processes that allow us to access the visual world from an experimental point of view. We need to enter our laboratories and test whether newly sighted subjects can positively pass Molyneux's test. However, the experimental question is underdetermined by a biological question: can we effectively restore vision subject born blind? The answer seems to depend on a set of very technical and complex factors about the nature of vision (Ferretti 2017b; 2019a).

In this respect, things are not very easy: visual perception depends on both the functioning of the eyes, as well as of the cortical visual brain. And restoring ocular processing could fail to be sufficient in order to have vision restored *tout court*: we should also be able to restore the cortical processing, at the computational level, involved in the manipulation of the visual information. However, we also know that, after a specific critical period of ontogenetic development, cortical vision cannot be successfully restored anymore, even if the eyes can absorb and process the light from the external environment (Gallagher 2005; Smith 2000; Ferretti 2017b, 2019a).

Several scholars have tackled the issue concerning the possibility of visual restoration at the cortical level, so as to permit to a subject's visual system to regain what is its proper function: visual recognition. However, since vision is a very complex and multi-layered phenomenon, asking about *whether it* can be restored leads us to ask *which aspects* of vision can be effectively restored. For example, in several subjects it is possible to restore the representations related to colour vision, but not other representations (shape or depth vision; see the cases discussed by Fine et al. 2003; Barrett e Bar 2009; Sacks 1995). Not only does neurobiology of vision help us in answering Molyneux's question. Furthermore, it is crucial to let us understand, from a conceptual point of view, what it means that a subject *can see* as well as whether it makes sense to ask whether a born blind subject can *come back to see again*. In this respect, it has been suggested that, if our cortical visual system is divided into two main pathways, the dorsal one involved in action guidance, and the ventral one involved in recognition (Sect. 2.1), we should then ask two different Molyneux's questions: one concerning the restoration of ventral visual processing, which would be the classic one concerning the recognition of the object *at first sight*; and one about the restoration of dorsal visual processing, concerning the possibility of an appropriate motor interaction with the object *at first sight* (Ferretti 2017b).

A specific look at the experimental results reporting the cases in which it has been tried to restore vision in congenital blind subjects, however, suggests the presence of several visual deficits in these 'Molyneux subjects'. Their visual recognition cannot be satisfyingly restored, so that they are in similar visual conditions to those patients affected by *visual agnosia*, whose visual recognition is impaired. Concerning the question about action, subjects are in visual conditions comparable to those patients affected by *optic ataxia*, in which visuomotor computations allowing to transform the visual stimuli in motor responses are impaired (Ferretti 2017b).

The conclusion is that, concerning the question about visual recognition, we cannot reach the scenario proposed by Molyneux's question: vision cannot be successfully restored. This does not allow us to positively answer the biological question, on which, however, the experimental one, about the effective test, seems to depend; if we cannot restore vision, we cannot ask about whether the subject would be able to recognize the shapes. Concerning the question about action, on the basis of our assumptions with respect to the story we endorse about the nature of the functioning of our visual system, either the answer is negative – the subject cannot successfully interact with the object *at first sight* – or the question is, as in the other case, without an answer, as we cannot reliably test it (for technical details, see Ferretti 2017b).

As for the previous section, the aim of this section was to show that philosophers have effectively taken into account the results from vision neuroscience in order to investigate

Molyneux's question, rather than providing a full apriori account. Recent accurate analyses are offered in (Ferretti 2017b, 2019a, Ferretti and Glenney, Under Contract).

The conceptual point at stake here is that, in order to answer the philosophical question about the nature of perceptual content, several philosophers have recruited the results from neuroscience: it is only by trying to understand how we can answer the biological question, as to reach the crucial experimental scenario at stake with Molyneux's question, that we can be in the condition of understanding the nature of cross-modal perception, which is at stake with the real spirit of the philosophical question.

To conclude, all we have said in the present and in the previous sub-sections clearly shows how philosophical questions - What is the difference between the real visual world and the pictorial visual world? What is the answer to Molyneux's question? - are investigated with a close inspection, and with specific philosophical analysis, of the experimental evidence from vision neuroscience.

3 From Philosophy to Neuroscience

Not only does neuroscience inform philosophy: there is also philosophy in neuroscience. In this respect, the need of rethinking the methods of neuroscience spurred a couple of debates that obtained the attention of philosophers of neuroscience.

It is quite agreed that a pivotal factor in the development of cognitive neuroscience was the availability of hemodynamic neuroimaging techniques: Positron Emission Tomography (PET), and, more relevantly, functional Magnetic Resonance Imaging (fMRI). Whereas the former makes use of a tracker, the latter is (usually) non-invasive, as it relies on an intrinsic physiological signal called *Blood Oxygen Level Dependent* (BOLD) (Cooper and Shallice 2010). BOLD signal exploits the difference in magnetic properties between oxygenated and deoxygenated hemoglobin to track neurovascular changes in specific areas of the brain. Such changes co-vary with differences in the neural activity of a given region (though this co-variation is not as straightforward as most scholars assume; see Logothetis 2008). The discovery of BOLD signal, thus, led to the possibility of measuring metabolic changes related to different cognitive tasks. This can lead to infer the corresponding neural activity for a given mental state: it is possible to assess which brain regions are *relevantly* more activated by some kind of cognitive process.

Whereas the enthusiasm about these techniques is well-deserved, it is worth stressing that neuroimaging scans are not *pictures* of the brain (Roskies 2007). However, even if we let asides the technicalities, some concerns remain on how to interpret neuroimaging data in cognitive terms.

The paper by the neuroscientist Richard Henson, *What can functional neuroimaging tell the experimental psychologist?* (2005), represents a remarkable endeavor to tackle this question by means of a formalization of two inferential models which have come to be known as *forward* and *reverse inference*. A forward inference obtains when a *qualitative* difference in the patterns of activation that accompany two behavioral tasks/conditions is taken as a proof that the tasks/conditions differ in at least one cognitive process. A reverse inference obtains when, during a given task/condition, the activation of some given brain structure is observed, and which was previously known to be involved in some cognitive process. In such a case, physiological

activation is taken as evidence that such a process is recruited in the context of the present task/condition. Let us briefly summarize some of the most relevant discussions strictly related to these two inferences.

3.1 Forward Inference and Bridge-Laws Connecting Psychology and Neuroscience

Let us begin with the epistemology of forward inference. In many respects, forward inference echoes the logic of dissociation in lesion studies (Davies 2010). However, the epistemic status of classical dissociations, based on lesion studies, is more robust than that of forward inference: in fact, while the impairment in some cognitive tasks, given by some brain lesions, counts as a rather direct evidence for a causal contribution of the damaged area, the activation portrayed by neuroimaging techniques might be merely epiphenomenal, i.e. it might reflect neural activity that regularly accompanies the task, but that is not necessary for it (cf. also Machery 2012). Painstakingly, it might be observed that even lesion studies cannot indisputably show that some area is *necessary* for a given cognitive process: in fact, in the brain (just as in many other biological complex systems) several functions are degenerate, that is, they can be realized by neural circuits that are at least partially distinct (Noppeney et al. 2004).

However, according to some researchers, not only is forward inference weak; it is a nonstarter. This critical stance is best expressed by Max Coltheart, an 'ultracognitivist' neuropsychologist who notoriously casted doubts over the psychological relevance of neuroimaging. In line with the functionalist metaphysics of the old-style cognitive science, Coltheart thinks that neuropsychology should aim at unraveling cognitive architecture, no matter how they are implemented.

In a target article that opens a forum on *Cortex*[4], Coltheart (2006a) implies the notion that neuroimaging does not shed light on mental processes. This notion is not defended with a frontal assault against forward inference. The strategy is rather that of offering a cunning challenge: namely, he invites his colleagues who are willing to play his game to exhibit some case study capable of clearly demonstrating "whether functional neuroimaging data *has already* been successfully used to distinguish between competing psychological theories" (Coltheart 2006a: 323). In the simplest case, two competing psychological theories T_a and T_b, imply that two cognitive tasks C_1 and C_2 are the product of a single cognitive system (T_a), or either of two distinct cognitive systems (T_b).

Coltheart (2006a) begins by reviewing the allegedly successful *forward inferences* discussed in the aforementioned paper by Henson (2005). In each case, Coltheart acknowledges that functional neuroimaging reveals the existence of a dissociation between two neural systems N_1 and N_2. However, he swiftly adds, this neural dissociation does not entail, *per se*, a cognitive dissociation: in his opinion, it might be the case that N_1 is distinct from N_2, and yet C_1 is one and the same with C_2; or vice versa.

The same appeal to underdetermination is invoked against all the case studies discussed by the many colleagues who participated to that forum. Coltheart (2006b)

[4] Vol. 42.

dismisses all their claims by following a similar plot: acknowledging that there is a dissociation between N_1 and N_2, but denying that this falsifies either T_a or T_b.

For instance, Umiltà (2006) discusses the following couple of theories:

(T_a) Endogenous attention and exogenous attention are governed by a single cognitive system C.

(T_b) Endogenous attention and exogenous attention are governed by two distinct cognitive systems C_1 and C_2.

The neuroimaging literature reviewed by Umiltà talks in favor of the existence of qualitatively distinguishable correlates for endogenous and exogenous attention, say N_1 and N_2. But Coltheart does not dispute this claim. However, Umiltà takes the dissociation of N1 and N2 as evidence in favour of Tb, whereas Coltheart denies it. At best, he claims, this dissociation can talk in favor of

(T^*_b) endogenous attention and exogenous attention are governed by a two distinct *brain* systems C_1 and C_2

over a rival hypothesis

(T^*_a) endogenous attention and exogenous attention are governed by a single *brain* system C.

It is worth stressing that Coltheart and Umiltà do not disagree about the nature of neural evidence. Their disagreement is, rather, at a different level: the level of the background assumptions. In the context of functional neuroimaging as a tool for psychological theory-testing, these assumptions concern mind-brain relationships (Nathan and Del Pinal 2016). While extremely simplified (or perhaps precisely in virtue of its simplicity), this debate offers a good spot for philosophers of science, for those scientific assumptions, usually lurking in the background of implicit reasoning, are brought to the surface and get formulated in explicit terms. That being said, as we will soon show, what scientists say (and think) they are doing sometimes does not correspond to what they actually do.

According to Henson, a forward inference such as that described by Umiltà only requires assuming a shallow structure-function mapping, which he calls *weak systematicity*: within the current experimental context, it cannot be the case that some regions are associated with a function in one condition, while other regions being associated with the same function in the other condition (Henson 2005: 215).

Truth be told, the requirements of Umiltà's forward inference are a bit stronger: in fact, such an inference is based on multiple experiments from different laboratories. Of course, generalization across different experimental contexts make the inferences more liable to mistakes: as we have mentioned above, some mental functions have been claimed to be degenerate, i.e. realized by distinct neural systems in different individuals, and perhaps by distinct neural systems even in the same individual across time, arguably due to the functional reorganization prompt by some brain lesions (Noppeney et al. 2004). That being said, those forward inferences that cannot be generalized outside a given experimental context will be of little interest. Because of this, Henson (2005) proposes to downplay the possibility of such phenomena of neuroplasticity by

positing that a normal function-structure mapping is the rule, whereas aberrant cases are treated as exceptions.

However, we do not need to delve in such sophisticated discussions to follow Coltheart's objection. As noted by Roskies (2009), while Coltheart affirms that he does "fully accept Henson's assumption that there is some systematic mapping from psychological function to brain structure" (2006a: 323), such a notion effectively happens to fail to draw the conclusions that follows from this assumption, for he *disallows* any inference from neural to cognitive dissociation. According to Coltheart, the only kind of data that can inform psychological theorizing are behavioral data (although he mitigated his position in later writings, switching the emphasis on the problem of underdetermination of neural data; see, for instance, Coltheart 2013).

While some principled reasons can be put forward in favor or against the bridge-laws underlying forward inference, Henson is probably right when arguing that this kind of assumptions can never be fully proven: instead, it is better to conceive them as 'theoretical bets' upon which scientific paradigms are built. Furthermore, rather than being demonstrated through *apriori* arguments, they should be assessed on the basis of the predictive and heuristic power they disclose.

3.2 Reverse Inference and the Pluripotentiality of Neural Structures

An even harsher, and thus unsettled, debate at stake with the destination of this part of or round-trip concerns the inferential model known as *reverse inference*. The notion of reverse inference has been formulated by Poldrack (2006), who refined the early definition of Henson (2005). The definition has the following structure:

(1) In the present study, when task comparison A was presented, brain area Z was active.
(2) In other studies, when cognitive process X was putatively engaged, then brain area Z was active.
(3) Thus, the activity of area Z in the present study demonstrates engagement of cognitive process X by task comparison A (Poldrack 2006: 59).

Poldrack warns his colleagues that mistaking this for a deductively valid inference implies committing the logical fallacy of *affirming the consequent*:

if P, then Q	*if it rains, there are clouds*
Q	*there are clouds*
Then P	*then it rains*

In order to be deductively valid, reverse inference would require that a given brain area Z is *selective* for a given cognitive process X, i.e. that it activates *if and only if* a cognitive process X is ongoing, but rests (relatively) inactive for every other process. In formal terms, it would require that premise (2) gets subsided by the stronger premise (2*):

(1) In the present study, when task comparison A was presented, brain area Z was active.
(2*) In other studies, *if and only if* a cognitive process X was putatively engaged, then brain area Z was active.
(3) Thus, the activity of area Z in the present study demonstrates engagement of cognitive process X by task comparison A.

The problem is that premise (2*) is likely false for every brain area (Anderson 2010). For instance, while Broca's Area has been historically related to language, its activation has been documented also during musical and motor tasks (Tettamanti and Weniger 2006). Therefore, in cognitive neuroscience there seems to be no room for such any attempt to straightforwardly deduce function from structure. This is hardly surprising: indeed, science rarely (if ever) advances thanks to *deductive inferences* – *pace* Popper. We need inductions, and arguably probabilistic tools to assess their soundness. Poldrack does not ignore that, however. Thus, after dismissing the deductive reading of reverse inference, he recasts it in probabilistic terms adopting the following Bayesian framework:

$$P(X|Z) = P(Z|X)P(X)/P(Z|X)P(X) + P(Z|\neg X)P(\neg X)$$

To exemplify the discussion, Poldrack makes use of the BrainMap[5] database, where the annotated results of hundreds of neuroimaging studies are stored. His query concerns the frequency of activation of Broca's Area (as defined by the coordinate of a previous study), in studies that involved a language task versus those that did not involve any language task (Table 1).

Assuming a prior P(X) = 0.5, that is, assuming that an unknown study has equal probabilities to either involve or not to involve a language task, the data presented in Table 1 warrants a posterior probability of 0.69. An appreciable gain, yet not a striking one.

Given that most areas seem to be pluripotent – that is, they are involved in multiple cognitive domains – how can we increase the predictive power of our reverse inferences?

To begin with, the alleged pluri-potentiality of neural areas can be interpreted in at least three (non-mutually exclusive) ways. Each of these ways is better dealt with a given cognitive strategy (McCaffrey 2015. See also Viola 2017; Viola and Zanin 2017).

First, the activation of an area in several distinct cognitive domains might be due to the compresence of distinct, but co-localized structures (e.g. neural populations) within it. Once isolated by specific tools and techniques, these neighbor structures might turn out to be far less pluripotent than they were initially thought to be. A major source of this problem is the smoothing of brain areas due to inter-subjective comparisons.

A second way to deal with pluripotentiality is by maintaining, on the one hand, that it actually occurs when one construes the workings of some area in behavioral terms, while insisting, on the other hand, that a more abstract functional characterization

[5] http://www.brainmap.org/.

Table 1. Experimental comparisons reporting Broca's Area being active or inactive in either language and non-language studies, retrieved from BrainMap (as of September 2005). Elaboration from Poldrack (2006).

	Language task (X)	Non-language task (¬X)
Total	869	2353
Broca's area active (Z)	166	199
Broca's area inactive (¬Z)	703	2154
Probability	P(Z|X) = 0.19	P(Z|¬X) = 0.08

would account for its (same) activity across several (distinct) cognitive domains. Coming back to the aforementioned case of Broca's Area, Tettamanti and Weniger (2006) have claimed that its cross-domain activity can be accommodated by the hypothesis that it plays a same cognitive role in all those tasks (language, music, and motion), to be construed at a more abstract level: namely, that of processing hierarchical structures.

Notice, however, that while this strategy can be applied in almost all cases, it is far from obvious that it brings more pros than cons. For instance, in order to accommodate the activity of the left posterior lateral fusiform across several apparently unrelated domains, Price and Friston (2005) redubbed it as a *sensorimotor integration* area. However, as stressed by Klein (2012), this kind of move is highly problematic: indeed, sensorimotor integration is such a vague functional ascription that it can apply to most parts of the cortex!

This brings us to a third way to deal with pluripotentiality: namely, coming to terms with it (rather than trying to 'negotiate it away'), and that we should live with it. This implies that we should do without (the hope to get) one-to-one mappings. This does not entail, however, that we should surrender to low predictive power. Instead, if the Broca's Area activity alone does not suffice to foretell if language or music is being processed, we might want to look at the activity of other brain areas. Indeed, given that no brain area produces a behavior in isolation, it seems reasonable to stop caring about which *areas* activate during some task, and rather care about the *networks*.

This is the logic underlying the now-prominent approach known as *MultiVariate Pattern Analysis* (Haxby, Connolly and Guntupalli 2014), which enables what popular science calls *brain reading*. By exploiting more powerful analyses, that allow to consider the activity of multiple brain regions at once, machine learning algorithms can be trained to correctly guess what kind of mental activity a subject is performing on the basis of its pattern of brain activation – and they are becoming increasingly good at doing so, even across subjects (i.e. when the neural pattern whose mental correlate they have to guess belong to a different subject than those whom which they have been trained). These kinds of predictions have been construed as *global reverse inferences* (Poldrack 2011; Nathan and Del Pinal 2017).

It is worth stressing that, despite their non-negligible predictive power, global reverse inferences do not straightforwardly inform us on what brain structures are causally responsible for some task. Aided by some examples, Ritchie, Kaplan and Klein (2019) have shown that the signal employed by the decoder to predict a mental

activity is not one and the same with the signal employed by the brain to realize it: it might be consistently produced, and yet causally inert. This is by no means surprising for those who are familiar with the epistemology of these tools: indeed, informative as they are, neuroimaging techniques are and will stay correlational in nature.

To sum up, in this section we have reviewed two inferential models underlying the scientific practice of cognitive neuroscientists: forward inference (brain dissociation implies a cognitive dissociation) and reverse inference (the activity of an area/networked involved in a given process is evidence for that process). While a careful exploration of the bundle of theoretical assumptions and problems underlying these models is obviously beyond the scope of the present discussion, hopefully they should have provided a rough idea of the kind (and the amount) of epistemological work to be done.

4 Conclusion. Philosophy and Neuroscience

In this paper, we have led the reader along a '*Roundtrip*' between philosophy and neuroscience. This *roundtrip* showed two important routes related to the contemporary reflection and the brain: on the one hand, theoretical, philosophical questions can be informed by empirical results from neuroscience; on the other, theoretical and epistemological, philosophical assumptions drive the research practices in neuroscience.

Though, for the sake of brevity, we decided to focus on four debates we are familiar with, plenty of similar cases can be invoked to witness such ongoing interactions. To name but a few: the debate on the relationship between action, intentions, and language (Ferretti and Zipoli Caiani 2018; Butterfill and Sinigaglia 2014; Burnston 2017); that about theories of emotions (e.g. Adolphs and Andler 2018; Celeghin et al. 2017); that about motor representations (Butterfill and Sinigaglia 2014; Ferretti 2016b, 2019c; Ferretti and Zipoli Caiani 2018a, Nanay 2013, Ferretti and Alai 2016).[6]

A major issue, when speaking of interdisciplinary interactions, is that of setting the boundaries of the disciplines. For instance, it has been objected, historical philosophical debates and modern scientific investigation over the Molyneux's question might stand in a relation of *natural evolution*, rather than involving a leap from one discipline to another.

We think that this kind of questions shed light over the nature of scientific disciplines and their boundaries in general. Rather than being *apriori* confined to some given layer of reality (in the sense of Oppenheim and Putnam 1958) or whichever domain of phenomena, we think that the (shallow) unity of disciplines is warranted by some (loose) methodological and theoretical commitments, as shown by the continuous hybridizations between disciplines (think for instance of economics' alleged imperialism toward other social sciences). If scientific specialization is mainly driven by historical and epistemic factors, rather than by ontological considerations, the presence and abundance of interdisciplinarity is unsurprising. On this view, interdisciplinary

[6] Cf. also the theory of lexical competence elaborated by Marconi (1997), originally meant to address some problem in philosophy of language, but subsequently tested in the scanner. (Marconi et al. 2013)

work boils down to the simple fact that scholars from different disciplinary traditions observe and study phenomena such as Molyneux's question from different angles within a *quinean* continuum of possible scientific perspectives.

In this respect, the facts disciplinary boundaries are first and foremost sociological (rather than ontological) does not imply that they can be trespassed without efforts or without risk. Scientific specialization poses serious challenges to interdisciplinary endeavors. On the hand, a reviewer for a philosophical journal who has to assess some claims based on neuroscientific evidence might not be in the best position to judge whether the author presented the findings fairly, or, conversely, cherry-picked those which best fit with her claims. On the other hand, a cursory look at the citation flows surrounding the two debates in philosophy of neuroscience shows that philosophers of science are better in importing scientific literature than they are in exporting their work to scientists – i.e. there are more philosophical papers citing scientific ones than the converse.

Disciplinary traditions and specializations arguably exist for good reasons – after all, nobody can be trained in everything. So, how can we satisfy the need for interdisciplinary studies without overlooking these reasons? Such questions were already framed by George Miller back in 1978. In his preface to the well-known Sloan Report on the state of the art of cognitive science, he acknowledged that:

> [a] revision of disciplinary boundaries seems called for, but such reforms are seldom successful; they have been likened to attempts to reorganize a graveyard. A more promising strategy is to recognize' the emergence of cognitive science by grafting new institutional structures onto existing ones. In several universities, collaboration has already begun among disciplines concerned with particular aspects of cognitive science. The most successful instances should be identified and encouraged to extend the scope of their activities. The particular institutional arrangements best suited to fostering such cooperation will, of course, depend on local custom, but it must be made possible to support directly the work of scientists whose interests fall between the traditional academic departments–the sort of scholar that each department wishes another would hire (ix).

Forty years later, his recipe – fostering interdisciplinary collaborations – still sounds like the best bet[7].

References

Adolphs R, Andler D (2018) Investigating emotions as functional states distinct from feelings. Emot Rev 10(3):191–201

Anderson ML (2010) Neural reuse: a fundamental organizational principle of the brain. Behav brain sci 33(4):245–266

[7] We wish to thank the audience of the 2017 Italian Association for Cognitive Science, as well as the audience of the 2017 Italian Society for Analytic Philosophy, for offering several good questions to an earlier draft of this project. We also want to thank those scholars who discussed with us about these topics: Silvano Zipoli Caiani, Giorgia Committeri, Bence Nanay, Andrea Borghini, Brian B. Glenney, Fabrizio Calzavarini, Gustavo Cevolani, Enzo Crupi. We also thank two anonymous reviewers for the comments.

Barrett LF, Bar LF (2009) See it with feeling: affective predictions during object perception. Philos Trans R Soc 364:1325–1334. https://doi.org/10.1098/rstb.2008.0312
Bechtel W (2009) Constructing a philosophy of science of cognitive science. Top Cogn Sci 1 (3):548–569
Block N (2014) Seeing-as in the light of vision science. Philos Phenomenological Res 89(3)
Brook A (2009) Introduction: philosophy in and philosophy of cognitive science. Top Cogn Sci 1 (2):216–230
Burnston DC (2017) Cognitive penetration and the cognition–perception interface. Synthese 194 (9):3645–3668
Butterfill SA, Sinigaglia C (2014) Intention and motor representation in purposive action: intention and motor representation in purposive action. Philos Phenomenological Res 88 (1):119–145
Celeghin A, Diano M, Bagnis A, Viola M, Tamietto M (2017) Basic emotions in human neuroscience: neuroimaging and beyond. Front Psychol 8:1432
Chinellato E, del Pobil AP (2016) The visual neuroscience of robotic grasping. achieving sensorimotor skills through dorsal-ventral stream integration. Springer, Switzerland
Coltheart M (2006a) What has functional neuroimaging told us about the mind (so far)? Cortex 42(3):323–331
Coltheart M (2006b) Perhaps functional neuroimaging has not told us anything about the mind (so far). Cortex 42(3):422–427
Coltheart M (2013) How can functional neuroimaging inform cognitive theories? Perspect Psychol Sci 8(1):98–103
Cooper RP, Shallice T (2010) Cognitive neuroscience: the troubled marriage of cognitive science and neuroscience. Top Cogn Sci 2(3):398–406
Davies M (2010) Double dissociation: understanding its role in cognitive neuropsychology. Mind Lang 25(5):500–540
Degenaar M, Lokhorst G-J (2014) Molyneux's problem. In: Zalta E (Ed) The stanford encyclopedia of philosophy. http://plato.stanford.edu/archives/spr2014/entries/molyneux-problem/
Dennett DC (2009) The part of cognitive science that is philosophy. Top Cogn Sci 1:231–236
Ferretti G (2016a) Pictures, action properties and motor related effects. Synthese 193(12):3787–3817
Ferretti G (2016b) Through the forest of motor representations. Conscious Cogn 43:177–196
Ferretti G (2016c) Visual feeling of presence. Pac Philos Q 99(S1):112–136
Ferretti G (2016d) Neurophysiological states and perceptual representations: the case of action properties detected by the ventro-dorsal visual stream. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology. studies in applied philosophy, epistemology and rational ethics. Springer, Cham, pp 179–203
Ferretti G, Glenney B (Under Contract) Molyneux's question and the history of philosophy. Routledge
Ferretti G, Alai M (2016) Enactivism, representations and canonical neurons. Argumenta 1 (2):195–217
Ferretti G (2017a) Pictures, emotions, and the dorsal/ventral account of picture perception. Rev Philos Psychol 8(3):595–616
Ferretti G (2017b) Two visual systems in molyneux subjects. Phenomenol Cogn Sci 17(4):643–679
Ferretti G (2017c) Are Pictures peculiar objects of perception? J Am Philos Assoc 3(3):372–393
Ferretti G (2018) The neural dynamics of seeing-in. Erkenntnis. https://doi.org/10.1007/s10670-018-0060-2

Ferretti G, Zipoli Caiani S (2018) Solving the interface problem without translation: the same format thesis. Pac Philos Q. https://doi.org/10.1111/papq.12243
Ferretti G (2019a) Molyneux's puzzle: philosophical, biological and experimental aspects of an open problem. Aphex, Open problems
Ferretti G (2019b) Perceiving surfaces (and what they depict). In: Glenney B, Silva JF (Eds) The senses and the history of philosophy, pp 308–322, Routledge
Ferretti G (2019c) Visual phenomenology versus visuomotor imagery: how can we be aware of action properties? Synthese https://doi.org/10.1007/s11229-019-02282-x
Ferretti G (Forthcoming). Why trompe l'oeils deceive our visual experience. J Aesthetics Art Criticism
Fine I, Wade AR, Brewer AA, May MG, Goodman DF, Boynton GM, Wndell BA, MacLeod DIA (2003) Long-term deprivation affects visual perception and cortex. Nat Neurosci 6:915–916. https://doi.org/10.1038/nn1102
Gallagher S (2005) How the body shapes the mind. Oxford University Press, New York
Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. Ann Rev Neurosci 37:435–456
Glenney B (2013) Philosophical problems, cluster concepts and the many lives of molyneux's question. Biol Philos 28(3):541–558. https://doi.org/10.1007/s10539-012-9355x
Gombrich E (1960) Art and illusion. Pantheon, New York
Henson R (2005) What can functional neuroimaging tell the experimental psychologist? Q J Exp Psychol Sect A 58(2):193–233
Jacomuzzi AC, Kobau P, Bruno N (2003) Molyneux's question redux.Phenomenol. Cogn Sci 2:255–280
Klein C (2012) Cognitive ontology and region-versus network-oriented analyses. Philos Sci 79 (5):952–960
Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453(7197):869
Machery E (2012) Dissociations in neuropsychology and cognitive neuroscience. Philos Sci 79 (4):490–518
Marconi D (1997) Lexical competence. MIT press, Cambridge
Marconi D, Manenti R, Catricala E, Della Rosa PA, Siri S, Cappa SF (2013) The neural substrates of inferential and referential semantic processing. Cortex 49(8):2055–2066
Matthen M (2005) Seeing, doing and knowing: a philosophical theory of sense perception. Oxford University Press, Oxford
McCaffrey JB (2015) The brain's heterogeneous functional landscape. Philos Sci 82(5):1010–1022
Miller GA (1978) Preface. In: Cognitive science, 1978. report of the state of the art committee to the advisors of the Alfred P. Sloan foundation. http://www.cbi.umn.edu/hostedpublications/pdf/CognitiveScience1978_OCR.pdf
Milner A, Goodale M (1995⁄2006) The visual brain in action, 2nd edn. Oxford University Press, Oxford
Nanay B (2011) Perceiving pictures. Phenomenol Cogn Sci 10:461–480
Nanay B (2013) Between perception and action. Oxford University Press, Oxford
Nanay B (2015) Trompe l'oeil and the dorsal/ventral account of picture perception. Rev Philos Psychol 6:181–197
Nathan MJ, Del Pinal G (2016) Mapping the mind: bridge laws and the psycho-neural interface. Synthese 193(2):637–657
Nathan MJ, Del Pinal G (2017) The future of cognitive neuroscience? reverse inference in focus. Philos Compass 12(7)
Noppeney U, Friston KJ, Price CJ (2004) Degenerate neuronal systems sustaining cognitive functions. J Anatomy 205(6):433–442

Oppenheim P, Putnam H (1958) Unity of science as a working hypothesis. Minn Stud Philos Sci 2:3–36
Price CJ, Friston KJ (2005) Functional ontologies for cognition: the systematic definition of structure and function. Cogn Neuropsychol 22(3–4):262–275
Poldrack RA (2006) Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci 10(2):59–63
Noë A (2004) Action in perception. The MIT Press, Cambridge
Poldrack RA (2011) Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72(5):692–697
Ritchie JB, Kaplan DM, Klein C (2019) Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. Br J Philos Sci 70(2):581–607
Roskies AL (2009) Brain-mind and structure-function relationships: A methodological response to coltheart. Philos Sci 76(5):927–939
Roskies AL (2007) Are neuroimages like photographs of the brain? Philos Sci 74(5):860–872
Sacks O (1995) An antropologist on mars: seven paradoxical tales. Knopf, New York
Siegel S (2010) The contents of visual experience. Oxford University Press, New York
Smith AD (2000) Space and sight. Mind 109(435):481–518
Schwenkler J (2013) Do things look the way they feel? Analysis 73(1):86–96
Tettamanti M, Weniger D (2006) Broca's area: a supramodal hierarchical processor? Cortex 42 (4):491–494
Tressoldi PE, Sella F, Coltheart M, Umilta C (2012) Using functional neuroimaging to test theories of cognition: a selective survey of studies from 2007 to 2011 as a contribution to the decade of the mind initiative. Cortex 48(9):1247–1250
Umiltà C (2006) Localization of cognitive functions in the brain does allow one to distinguish between psychological theories. Cortex 42(3):399–401
Viola M (2017) Carving mind at brain's joints. the debate on cognitive ontology. Phenomenol Mind 12:162–172
Viola M, Zanin E (2017) The standard ontological framework of cognitive neuroscience: some lessons from Broca's area. Philos Psychol 30(7):945–969
Wollheim R (1998) On pictorial representation. J Aesthetics Art Criticism 56:217–226
Zipoli Caiani S, Ferretti G (2017) Semantic and pragmatic integration in vision for action. Conscious Cogn 48:40–54
Zipoli Caiani S (2013) Extending the notion of affordance. Phenomenol Cogn Sci 13:275–293

Abduction, Problem Solving, and Practical Reasoning

The Dialogic Nature of Semiotic Tools in Facilitating Conscious Thought: Peirce's and Vygotskii's Models

Donna E. West(✉)

State University of New York at Cortland, Cortland, USA
westsimon@twcny.rr.com

Abstract. Peirce's adherence to the endoporeutic principle reveals how abductive rationality can effectively be exploited; it elevates dialogue as the most efficacious intervention to advance the development of modal logic. Peirce's endoporeutic principle prefigures Vygotskii's private and inner speech as the primary factor affecting thought refinement. This inquiry explores the far-reaching effects of internal dialogue upon hypothesis-making, particularly critical at early ages. In short, Propositions expressed in talking to self may obviate which hunches surface in their infancy as faulty/plausible, in a way that no other kind of interventions can insinuate, creating what Peirce describes as "double consciousness." Double consciousness privileges the element of surprise within dialogic exchanges (linguistic, and non-linguistic alike) by means of the imposition of the "strange intruding" idea. Double consciousness as self-talk can preclude adopting emergent propositions/assertions (often only implicit), it offers a powerful forum to discard hasty/weak hunches in a timely fashion, together with those whose content fails to give rise to serviceable courses of action and workable remedies.

1 Introduction

Peirce's and Vygotskii's accounts of the dialogic nature of conscious thought follow rather distinctive paths – the former recommends indexical shifts via double consciousness, while the latter advocates use of more symbolic linguistic systems to create the conflict of motives necessary to problem-solve. Whereas both models advocate the necessity for dialogically-based tools to advance logical reasoning, they suggest quite distinctive kinds of representations (mental tools) to advance conscious recognition of legitimate perspectives. Vygotskii and Peirce both recognize that to be efficacious, any intervention advancing dialogic reasoning must first identify particular ends, and must construct a feasible plan to reach such ends. The plan/course of action needs to specify the participants necessary to the success of the plan, and must identify the psychological, social, and logical consequences likely to proceed therefrom. While Vygotskii is more explicit in identifying tools and their direct effects, he (unlike Peirce) often fails to make mention of indirect effects – which may overlook the meanings underlying tools. In Peirce's semiotic meaning is not separate from the tool as a sign, but is included in the sign itself.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 193–216, 2019.
https://doi.org/10.1007/978-3-030-32722-4_12

Peirce's semiotic model offers the benefit of extracting meanings, since sign meanings are present upon the existence of the sign, prior to their implementation (cf. Deely 2009 and Deely 2012). As such, these embryonic or potential meanings are influential not merely on the synchronic plane, but have a diachronic, amplified effect – stimulating science, and aesthetic endeavors throughout generations. Although the tools which Vygotskii identifies have more clear practical effects toward practical ends [implicating the development and utilization of practical abductions or as Magnani (2017: 17) terms them "instinctive abductions… in practical reasoning"], the effects of the tools that Peirce advocates (although practical (manipulative as Magnani 2018: 147) terms them, precipitate more objective ends. His primary tool to promote conscious logical reasoning is, in fact, abduction, deriving either from flashes of insight which ultimately benefit the continuum or from more deliberated insight (cf. West 2015). In contrast, Vygotskii's model emphasizes the influence of action upon self/another (either memory aids or responses of a more mature thinker) which promote the contemporaneous labor effort.

Conversely, Peirce's semiotic privileges tools whose truth value has a more objective more historical end; it places front and center truth beyond the good of the society. Its intent is to demonstrate how well-founded hunches can inform and reform the perspectives of all mankind, scientifically, emotionally, ethically, aesthetically, and otherwise – motivated by intra and intersubjective dialogic exchanges. Although these hunches often begin as sudden glimpses within ego alone, as in children's early percepts illustrated in virtual habit (Bergman 2016 and West 2017), and are generated upon scant evidence by a single individual, they offer some nugget of viability to uncover the Final Interpretant: in human ontogeny and for the continuum at large "there is certainly a third kind of Interpretant, which I call the Final Interpretant, because it is that which *would* finally be decided to be the true interpretation if consideration of the matter were carried so far that an ultimate opinion were reached" (EP 2: 496). As such, they qualify as tools for the advancement of the state of logic for the entire continuum; and their meanings/effects extend beyond human applications.

This inquiry will provide evidence that the most effective tools to cultivate inferential reasoning are those which hasten assumption of other perspectives in speaker and listener reciprocal roles. These kinds of tools (which promote intra and intersubjective dialogue) especially help children to advance meanings/effects legitimately attaching to particular events/episodes. It will be demonstrated that switching from agent to receiver roles unquestionably stimulate both practical and intellectual abductions for the self and for the entire semiosphere – be they Vygotskian or Peircean. Their semiotic character as acquired by children and as expressed by Peirce – triadic relations of: representamen, object and meaning/effect (c. 1896: 1.480)[1], affords increased means in the course of development to recognize the need for modifications to originary sign associations, because with dialogically communicated sign-meaning pairing, signs begin to represent something other than themselves; and their foundation is, by nature, teleological

[1] "… representation necessarily involves a genuine triad. For it involves a sign, or representamen, of some kind, outward or inward, mediating between an object and an interpreting thought. Now this is neither a matter of fact, since thought is general, nor is it a matter of law, since thought is living."

(Gallagher 2017: 88). In that the sign-object-meaning associations that children make insinuate particular "functions toward x," they reveal some foundational recognition for the relation's purpose.

Peirce's inclusion of meaning has implications as to purpose, because his rationale for inclusion of meaning in the sign itself (as a necessary and separate feature) was to preclude misrepresentations that might arise between representamen and the intended meaning, especially surfacing in the dialogic arena (Gallagher 2017: 88). Peirce's emphasis upon sign-object-meaning in semiosis obviates the importance of augmenting meaning of the same sign over time; and decoupling meaning/objects from X (cf. Leslie 1987, and West 2013), and recombining structures and meanings into different representational frameworks are facilitated (cf. Gallagher 2017: 88). Hence, Dialogic exchange of signs which suggest alternative object/meaning associations carry great weight in convincing one to augment originary concepts. Sharing with the self in private or inner speech, or sharing with another in social speech, constitutes effective tools to establish and execute new propositions/arguments – by suggesting new relations between objects and events. These dialogic tools (sharing novel semiotic meanings) constitute auxiliary signs, and provide children external vehicles to search out latent relations between contributory events and consequences. For Peirce and Vygotskii, sharing diverse viewpoints (on the internal logical or discourse plane) extend tool use, with their means to initiate alternative modes of consciousness. The consciousness that perspective exchanges hasten communication of myriad glimpses into relational logic.

Without change in belief or action that novel meanings afford (cf. West 2016 chapter 13, and Kilpinen 2016), Peirce asserts that the tools used to orchestrate advancements are ineffectual, and must be reformed, despite their value in determining what is not true (1906: 4.538). This underlies the necessity (which Peirce recognized) for widening certain signs to incorporate additional interpretants toward the business of refining logical event relations. These logical relations (taken up in Peirce's later works) are crucial to dialogic forms of meaning-making; they elevate indexical and iconic tools as primary instruments in the business of exchanging new perspectives/meaning relations.

2 Foundations of Peirce's Dialogic Inferencing

Unquestionably, dialogue represents one of Peirce's primary tools to invoke habit-change. It can do so through narration –scaffolding event organization onto discourse units, as Bamberg (1992, 1997) proposes. Or intentional behavior schemes surface constituting "online feedback modulated adjustments …below the level of intention, but collectively promote the satisfaction of an antecedent intention" (Rowlands 2006: 103). In other words, tools emerge prelinguistically, prior to dialogic exchanges, and are later motivated by dialogic interactions. Tools which obviate early event schemas on a non-dialogic level are indexes; and the organization of events –integrating sign, object and function/purpose, is first illustrated by index via the Peircean category of Secondness. Even at early ages, gesture illustrates and hastens event punctuality/telicity and progressivity, at the prelinguistic level. Prelinguistic gestures (targeted reach, maintaining attention to objects via gaze/pointing) exemplify resistance or effort

against a new force. Such constitute indexical tools to individuate objects/places in the here and now and to measure their imposition (effort, resistance) upon the organism. Early uses of index (in developing gaze trajectories and prehensile skills) obviate index's indispensable tool-based function (cf. Bamberg 1997; West 2013, 2016, 2019, under review). Given its early intervention-based function as gestural organizers on the physical plane, and its protracted utility as viewpoint regulators on the mental plane, index's role as tool to advance inferential thought is legendary. In its latter function, as internal tool, Index serves as the catalyst for the advent and refinement of dialogue as narrative. Despite their status as Seme (1903: EP2:274), gestural indices can imply concepts, actions, and assertions, when they acquire the status of Seme. The progression is as follows. When physical gestures constitute the platform to individuate events, and consolidate them into episodic units, their interpretants are expanded. Here index suggests what children should do in particular situations (Pinto et al. 2016; Pinto et al. 2018; Trabasso and Stein 1997). It draws an attentional template between actors, objects and goals in events; it promotes give-and-take exchanges between interlocutors. These exchanges entail looking toward another and directing participants' gaze trajectories toward an individuated object.

Index affords this orchestration of shared dialogue-building by virtue of its natural means to obviate joint attentional contours, in turn enriching the meanings which interlocutors attach. This use of index emerges at 1;3 (Saylor 2004; Baldwin and Saylor 2005). This joint directional medium is responsible for advancing attentional enterprises from individual to social, a precursor to dialogue/narrative. In short, exploiting indexical meanings within viewpoint genres constitutes the most effective tool to learn about shifting theories of mind critical in dialogue-building. According to Peirce, successfully tracing viewpoints enhances interpretation in the form of establishing "common ground" or a "place to stand" shared by participants (1906: MS 614). This emphasis on the part of Peirce to establish a shared focus on the same objects and meanings in the outer world (as world knowledge) is pivotal to the exchange of signs; otherwise, interpretability would be immaterial to sign use. Pietarinen (2006) elaborates on Peirce's reliance upon common ground issues, capitalizing on what he terms Peirce's endoporeutic principle. In fact, Peirce advocates that establishing this "common ground" or "place to stand" is an indispensable step to "know the universe" (1902: 3.621).

3 Early Tools Measuring Dialogic Communication

Gestural performatives (gaze, pointing, arm extension), whose onset is simultaneous to joint attentional gestures constitute indexical tools stimulating dialogic communication. They likewise depend heavily upon the attention-directing power of indices. Performatives are operational in that with intention (cf. Austin 1962) the attention of an interlocutor is shifted to an individuated object (cf. West under review). The performative (social exchange) takes the form of an indexical action (pointing gesture with motion toward X) whose intent is to communicate either a declarative or an imperative – the intent that another do something. The interpretants of these gestural performatives are enhanced when holophrastic and telegraphic utterances emerge at approximately

1;6 (Clark 2009); and with the emergence of double indices (in gesture and language) the propositions underlying interpretants of index become more explicit, less implicit, such that their meaning is disambiguated. For this reason, they more clearly qualify as Phemes,[2] particularly consequent to the use of two indexical signs to communicate the same interpretant. As such, in Peirce's later semiotic (1906), he accounts for propositions that were merely implied (cf. Bellucci 2014: 539–540 for elaboration). In Peirce's new taxonomy of the dicisign, he uses Semes to imply propositions in Terms, and uses Phemes to imply propositions/arguments in actions with two quite different indices, while maintaining the meaning. The greater attentional force of the latter qualifies indexical gestures as Phemes. The Pheme is tantamount to an imperative or a compulsive act (1906: MS 295), e.g., pointing accompanied by a demonstrative (cf. West 2013). In fact, Peirce explicitly characterizes the Pheme as an index (1906: MS 295: 26), in that it often gives rise to an immediate response, similar to the effect of an action performative. In MS 295: 26, Peirce provides examples of Phemes (as acts of nature or actions compelling automatic conduct) which clearly showcase effects produced by performatives, namely, an earthquake, or simply a sudden action, as in a call to arms (1907: MS 318). Peirce does not restrict his use of Pheme to action-based commands/performatives, but extends its use to illocutions/perlocutions—since language-based commands likewise qualify as Pheme. The latter is obviated in MS 295 when Peirce characterizes the Pheme: "intends or has the air of intending to force some idea…upon the interpreter…." By "interpreter" Peirce refers not merely to existing interpreters, but to all possible interpreters within the continuum—past, present, and future. The force of the Pheme rises to illocutionary and perlocutionary status in that implied promises across agents unfamiliar to one another are proper.

Early on, before indexical gestures illustrate motion or joint meanings, they are pre-performative, and qualify as Semes (West under review). When they embody movement trajectories and joint responses possessing a performative character, they acquire the status of Phemes, in that their effect causes compliance with an established code of conduct or conceptual standard, e.g., dropping an object on the floor to force pick-up. Peirce describes the intentional and brute effect (on another's response) of the Pheme as follows: "Such a sign intends or has the air of intending to force some idea (in an interrogation), or some action (in a command), or some belief (in an assertion), upon the interpreter of it, just as if it were the direct and unmodified effect of that which it represents" (1906: MS 295). The "direct and unmodified effect" is so integrally connected with the representation that the response is verbatim, without conscious deliberation. To illustrate, the dynamic interpretant (direct effect) translates into an automatic response – a response implemented by virtually all members of that society. This universality of response further demonstrates the brute power of the indexical sign, such that pick-up of a child's discarded object or retrieving arms for battle, flow despite alterations in the context in which the sign materializes.

[2] "The second member of the triplet, the "Pheme," embraces all capital propositions; but not only capital propositions, but also capital interrogations and commands…. Such a sign intends or has the air of intending to force some idea (in an interrogation), or some action (in a command), or some belief (in an assertion), upon the interpreter of it, just as if it were the direct and unmodified effect of that which it represents" (1906: MS 295: 26).

Similarly, more advanced syntactic utterances accompanied by moving gestural indices still qualify as Phemes, because they give rise to universal responses from interlocutors. Despite the formulaic paradigm of these more advanced utterances (ready-made strings), they (like static indexical gestures) call for brute responses, and likewise qualify as Phemes. The most basic of narratives emerge at this juncture, namely, nursery rhymes, which, although they resemble stories with beginnings, middles, and ends, are characterized as merely a verbatim series of events which are virtually never altered (Fivush and Haden 1997). As such, they are not subject to change/modification – incorporating novel beginnings or substituting alternative conclusions. For this reason, nursery rhymes do not refer to past or future events, and do not constitute full-fledged narratives.

Other kinds of static, unaltered episodes surface at the same age, between 1;0 and 3;0, namely, enactments (McNeill 1992; Acredolo and Goodwyn 1990; cf. West under review for further elaboration), and flashbulb memories (Neisser 2004). The former materialize as action sequences bounded by the same purpose; and their structure is often more complex than is children's syntax at the same age (Stanfield et al. 2014). The latter (flash bulb memories) feature different types of personal experiences: a hospital stay, or relocation to a different residence (cf. Neisser 2004). Like nursery rhymes (given that their story-line always proceeds in a certain sequence), flashbulb memories constitute signs which remain virtually unalterable, because their parts are not perceived to be separate units, at least at the outset; these kinds of memory likewise elicit the same effects – identical emotive and action responses. They ordinarily consist in recall of single vivid picture-based personal experiences which feature a particularly sensational event (either affirmative or negative); and the autobiographical nature of these memories, in turn, contributes to their resistance against alteration of the sign and its interpretant.

The next sign which serves as a tool for advancing association of different meanings to the same sign is Vygotskii's notion of private speech. There are certain stages and it proceeds dialogue to others, just that to initiate an event pulls it through the goal, to its resolution. Whispering is the second stage of private speech. It becomes what Vygotskii refers to as inner speech, wherein audible articulation of a path to a solution is unnecessary to reach the goal. Between 3;0 and 4;0, audible self-talk serves as the transition from formulaic cause-effect event associations to more reasoned ones, such that logic governs the connections between antecedents and their consequents. Children generate audible sequences of utterances in the process of constructing how they will address their actions to problem-solve, e.g., "now I am going to do X, then X, and afterward X." The process of using speech to enhance goal directed event logic begins with organizing precedent and resultative events according to the degree of effect upon the consequence ordinarily to the self are scaffolded onto event schemas, such that the need to utilize a procedure is vitiated in order to regulate ourselves—the critical component of Peirce's notion of mature thinking is self-regulation. Findings from three and four-year-olds (Winsler et al. 1997: 75), are in accord with this premise: "These findings suggest that the movement from interpersonal collaboration to independent problem-solving involves children's active participation in taking over the regulating role of the adult collaborator. The suggestion here is that in the development

of cognitive functions children use private speech to collaborate with themselves in much the same way that adults collaborate with children during joint problem solving."

Certain linguistic collaborations are especially instrumental in regulating self as agent – those which demand double viewpoints, because they require children's assumption of shifting roles necessary to articulate private speech. Increased use of listener pronouns and Telephone interactions are particularly useful to this end. Telephone interactions are especially efficacious in children's transition from private speech to inner speech – audible self talk to inaudible self-talk, since it requires children to reflect upon and make assumptions regarding the state of others' knowledge (epistemic assumptions), and their emotional inclinations toward the subject of discourse (deontic assumptions). Additionally, speakers must realize that access to the physical context afforded to them is not afforded to the listener; and as such, spatial coordinates critical to the subject of discourse must be made explicit, e.g., sufficiently describing depictions in a story book.

To establish "common ground" (joint focus), speakers need to appreciate that it is they who have the burden of determining what must be made explicit, what implicit, and what information can be omitted. These decisions hinge upon the accuracy of speakers' theory of mind assumptions. Inadequate preparation to this end, or simply lacking knowledge of the listener's unique epistemic or deontic base can depress strides toward establishing and maintaining "common ground." Moreover, speakers' misplaced assumptions can likewise depress apprehension of interlocutor's different logical connections – those representing their episodic assumptions. If the speaker's assumption is that a listener already recognizes and validates a particular event relation, supplying detailed explanations as to its plausibility is unnecessary; and inclusion of extraneous (known information) can be rather confounding.

Several researchers, among them, Cameron and Lee (1997), Cameron and Wang (1999), and Cameron and Hutchison (2009) examined children's linguistic adaptations when using a telephone as a communicative device. Their purpose was not merely to measure the kinds of modifications in the telephone medium (compared with face-to-face interaction), but to monitor the success of the device in hastening private speech. At 3;0 and 5;0, children were instructed to tell a story to a familiar interlocutor (*Frog, Where are You?*) in two conditions: over the telephone and face-to-face. The book displays a series of pictures about a boy and his dog attempting to capture a particular young frog who escaped from a jar while at their home (Mayer 1969). The picture sequence demonstrates the locations (where the characters explored) to orchestrate the capture and in what sequence. Children were expected to describe and explain the happenings as per narrative structure. Findings illustrated clear differences between the telephone and the face-to-face conditions. In the telephone phase, subjects' utterances were more elaborated – their utterances contained more descriptions, and were longer (Cameron and Wang 1999). Subjects' increased need to identify and situate objects and events for the addressee while on the telephone contributed to decreased use of gestural pointers, demonstrative pronouns, and personal pronouns), but greater dependence upon iconic descriptions and informational indices (cf. Stjernfelt 2014 for elaboration) to paint mental images for the listener. The use of informational indices disambiguates who does what to whom, and where and when (cf. West 2013; West 2018), which forces children to make explicit these factors for themselves in private speech.

Alderson-Day and Fernyhough (2015: 939) suggest another intervention – referring to one's self as "you." This intervention is likewise effective in languages which mark listener roles with affixes or with pragmatic cues, because practice assuming addressee viewpoints is still operational. This strategy forces less experienced speakers to reflect upon and assume listener roles for themselves, in turn, significantly facilitating private speech, and promoting improved problem-solving skills. This same approach has been advocated as an intervention for children at 3;0 (cf. West 2011). The rationale is as follows: referring to the self as "you" highlights the inherently shifting character of speaker-listener roles so necessary to constructing private speech, in that it forces greater objectivity in viewpoint paradigms. In short, narrating events into episodes via the medium of listener role or via the telephone forces children to spatially and temporally situate constituent events for themselves and for others. In this way, episodic features can be made more prominent, in turn enhancing apprehension of logical relations across individuated events.

With these kinds of linguistic tools, children become proficient at perceiving and expressing the refined event relations inherent in inner speech, and mature logical systems (cf. Cameron and Wang 1999 and West 2014). As a consequence, arguments which were once implicit, are expressed explicitly, but are likely to obviate effects on the self. As such, argument structure at the three-year mark organizes locations but not time coordinates. While at 3;0 children can make explicit place sequences, where objects have been transferred to a hiding place (Hayne and Imuta 2011), their competencies at 4;0 are far more elaborated – extending to expression of temporal sequencing (Tulving 2005). At 4;0 children are competent at narrating not merely where past events took place, but when events materialized with respect to one another (Perner and Ruffman 1995). When telling about a fire alarm incident, children (at 4;0) were accurate at narrating the places and times of past events which they, themselves experienced (Pillemer and White 1989). The temporal organization, however, pertains to past events only; and those events were autobiographical in nature.

When children begin narrating to themselves inaudibly (as in Vygotskii's notion of inner speech), they need to depend upon syntactic competencies intrinsic to articulation; hence their working memory resources are freed-up to generate semantic and logical relations. Likewise, when engaging in inner speech, children take double albeit not conflicting roles – as generator of utterances, and as considerer of the viability of those utterances (reflecting upon their logical promise). Fernyhough (2008: 239) supports the fact that inner speech is a form of Peirce's "double consciousness," which advances the generation of promising logical relations, in that inner speech "…provide [s] a link between intentional agent and mental-agent understanding."

4 Speech to Consciousness in Vygotskii's Paradigm

Vygotskii afforded speech the most primary role in transitioning from unconscious to conscious thought, beginning with social speech, proceeding to egocentric speech (private speech), and finally to inner speech. Private speech is entirely audible; it is uttered by children to themselves, often to develop a plan of action for problem-solving. Its form follows the syntactic conventions relevant to the particular language,

such that both the sequence of words within sentences/clauses and the inclusion of lexical items are overt. In fact, private speech constitutes a tool to enhance self-talk and to construct strategies for planning activities. It does so by replacing dependence upon adult input with children's own intra-collaboration skills: "The movement from interpersonal collaboration to interdependent problem-solving involves children's active participation in taking over the regulating role of the adult collaborator. The suggestion here is that, in the development of cognitive functions, children use private speech to collaborate with themselves in much the same way that adults collaborate with children during joint problem-solving" (Winsler et al. 1997: 75). In audibly articulating to themselves, children share with themselves a dialogic perspective – they take both speaker and listener roles, and reverse them to test the viability of the individual perspectives.

In the transition to inner speech the order of the lexical items and their inclusion becomes altered. In many cases, it becomes difficult for interlocutors to comprehend intended meanings of child speakers when syntax is truncated and words are deleted en route to inner speech. This incomprehensibility likewise surfaces in self-talk, when because of lexical omissions, children are more likely to forget what they are saying to themselves. According to van der Veer and Valsiner 1991: 366, when private speech is en route to being internalized, it acquires a "special syntax," characterized by fragmentation, abbreviation, and with "a tendency towards predicativity." The latter results in "a tendency towards omitting the subject of the utterance" (van der Veer and Valsiner 1991: 366). In fact, omitting subjects places grater dependency upon the linguistic components which are remaining – none other than predicates. For example, deleting the subject from "the chair is blue," leaves us with the predicate "is blue," placing the emphasis upon the quality of blueness for the already established (omitted) subject. This truncation can be advantageous or disadvantageous. It allows less constraint in the free-flow in the characterization of entities and their effects; at the same time, it can contribute to gaps in understanding the proposition, if the interlocutor fails to hear/forgets the subject.

Nonetheless, neither van der Veer, et al. nor Vygotskii himself ever discuss rationale for why or when subjects are deleted during this developmental stage; but, children's spontaneous productions and their elicited imitations of adult utterances evidence subject omission well before 4;0 when inner speech becomes the practice. This phenomenon of subject deletion or null subject particularly surfaces when the subject is pronominal (Montrul 2004: 190; Hyams 1994; Valian 1991; Valian and Aubry 2005). The omission of pronouns in subject and sentence initial position illustrates unconscious but deliberate deletion of less critical information – old information (or that which is not the topic of discourse). So, approximations toward inner speech (truncated forms) can reveal children's semantic and pragmatic presumptions – what does and does not need to be repeated to the listener, and demonstrates which information is assumed to be more or less important for intersubjective interpretation.

In characterizing the linguistic processes to transition from private to inner speech, Vygotskii intimates a critical distinction between thought and language – that thought consists in proto-propositional cognitions which are unconscious, while, consequent to their structural constraints and omissions of such, private and inner speech rise to the level of conscious mental processes. The latter is so, given the special nature of

language systems – requiring sequential structure and overt representations. The explicit and overt nature of words allows interlocutors to consciously attend to components of the proposition to be communicated, and establishes organized problem-solving behaviors. This tool-like function is orchestrated when symbolic signs (words) force the producer to make the proposition intelligible to an interlocutor. Hence, the more conscious the sign's interpretation, the greater is its tool-like effect to enhance dialogic interpretation.

In fact, the overt nature of private speech privileges it as a tool to facilitate inner speech (internal speech) – ultimately regulating both intersubjective and intrasubjective communication. Because private speech expresses meanings (in talk to and from self) via overt, linguistic signs, its tool-like effect – to display signs with conventional, reliable interpretations is more likely. The fact that overt speech has a measure of syntax, as well as agreed upon semantic import demonstrates its means to organize the rather amorphous character of thought toward achieving an internalized, but reliable (conventional) message. For this reason, private speech is a necessary tool to ascertain inner speech and to regulate unstructured thought, because the organizing structure of its syntax exploits the substance of propositions, as well as their boundaries. Hence, private speech supplies the power to transform fleeting propositions within thought to arrive at a more constrained clausal structure still required (although unexpressed) in inner speech.

To elaborate, private and inner speech supply the power to heighten propositional potency via their overt/covert syntactic elements. Adherence to structural constraints constitutes a force which individuates propositions as single claims. But, at the same time, the implementation of lexemes in private and inner speech compels recognition of logical relations across propositions, providing the raw material to fashion arguments. Propositions are more realizable when clausal boundaries separate one proposition from another, and, more importantly, when they suggest the nature of their logical relations by means of subordinate clausal structure (cf. Lust et al. 1987). Lust et al. (1987) demonstrate that at 3;0 and beyond, children have a tendency to alter sentences which they are asked to repeat, such that imbedded clauses (subordinate clauses) are converted to coordinate clauses, e.g., adult: "because the sun is bright, the mother is sneezing," child: "the sun is bright and the mother is sneezing." This tendency on the part of children undoes the argumentative structure of the adult's utterance, only maintaining its internal but perhaps unrelated, propositional value.

Goudena (1992: 216–217) suggests that the tool which facilitates conscious consideration of different perspectives is, in fact, syntax, because it directly enhances the productivity of propositions by suggesting the nature of relations between events. He intimates that problem-solving with private speech facilitates consciousness of contextual features critical to connecting contributory events to their consequences, obviating spatial and temporal referent points across events.

Nonetheless, the transition from private to inner speech requires more than the implementation of syntax onto indescrete forms of thought. Once consciousness is in place consequent to overt speech, syntax can play a diminished role in proposition-making. As such, truncation (omitting sentential subjects), and the implementation of paralinguistic tools (intonation) obviate the semantic and pragmatic determinants which underlie propositions (cf. Goudena 1992: 216–217; and van der Veer and Valsiner

1991: 367 for an alternative discussion). Syntactic omissions which usher in inner speech ordinarily include subjects, especially those which represent information already mentioned in the discourse, or already within the stored knowledge of both interlocutors (consequent to shared experience or shared interests). This pragmatic competency depends, in large part, upon recognition of topic shifting in narration, or a determination of the psychological subject. As Goudena observes: "…the psychological subject need not be mentioned; speakers know what they are talking about" (1992: 216). The fact that pronouns as subjects are likely to be both deleted and non-topicalized in sentence initial position supports the latter claim, because pronouns are more likely than are nouns to refer to old information (Halliday and Hasan 1976), and are more likely to be deleted items in the truncation process (West 2011 and West 2013). Alterations in intonation contours are instrumental in differentiating whether the speaker intends to make an assertion (declarative); or whether the intent is to convey a command/recommendation to the listener.

5 Vygotskii's Use of Memory Tools as Semiotic Devices

Peirce's semiotic can inform the use of Vygotskii's memory tools to enhance conscious problem-solving. To illustrate, Vygotskii's memory tools (his double stimulation paradigms) incorporate Peirce's indexical sign. In fact, his paradigms utilize indexical pointers to advance children's conscious memory to transition from physical signs to internal ones to facilitate memory. Consequently, Vygotskii's intervention strategies supersede psychological instruments to advance from sensory-motor to logical intelligence; they represent a semiotic measure, in that what aids memory of past decisions and organizes selection of subsequent steps toward resolution are indexical signs. These indexical memory devices mark past steps from subsequent ones, and force focus upon future-thinking – upon the eventual consequences likely to materialize within problem-solving paradigms. In support of this claim, Sannino (2015: 9) asserts that, like a double stimulation memory tool, "the pointing finger is "the key component in mastering mediated attention." In fact, pointing, as well as other indexical signs control children's attention at early stages in ontogeny, and eventually establish conscious noticing are pillars in demarcating spatial and temporal boundaries. In this way, they allow children to exploit memory of previous actions from what still needs to be accomplished. Indexes are used universally at primary stages in human ontogeny to individuate spaces, actions, and events, and to make salient relations across objects and events (cf. West 2014).

Because these semiotic devices direct attention to physical objects, and hasten conscious notice of logical relations between objects/events (cf. West 2013 and West 2019), they occupy a particularly facilitative role in Vygotskii's double stimulation paradigms, e.g., pointing hands on a watch to illustrate that it is time to arrive at a decision to act. These double stimulation memory tools control children's attention to and conscious notice of problem-solving strategies, such that physical stimuli serve as reminders of steps already engaged in/prospective steps toward resolution. At this juncture, children's attention is fastened upon a physical representation which induces them to implement the next step to resolve the problem. The latter stimulus illustrates

that children recognize the possibility of particular consequences in the face of action/non-action.

The primary impetus for Vygotskii's (derived from Leont'ev) method was to create a kind of forced decision-making paradigm, such that isolating presentation of relevant stimuli or their qualities reminds children of already achieved steps in the problem-solving effort. At the youngest ages (4;0), Vygotskii's initial paradigm did not utilize physical aids, but relied upon children's memory alone. After instruction (in the process of making certain color associations), children were expected to respond rapidly to questions regarding the color of objects in particular contexts. The instructions constrained children to refrain from "yes"/"no" responses, or to avoid use of certain colors (beyond initial use). Sample questions included: "do you go to school?" "were you ever in the hospital?" Inclusion of physical signs which are indexical in nature (colored cards) were introduced to further enhance memory (beyond the verbal instructions). These colored cards became place markers to hasten memory of the colors which were already employed. Children were expected to assign less typical qualities and purposes to familiar stimuli. Vygotskii refers to this testing paradigm as "double stimulation," applying it to children beginning at 4;0. The impetus for this intervention was to compel subjects to draw upon conscious means to dissociate conventional, automatic sign-object-meaning connections, and instead to substitute less obvious qualities/purposes to entities/events.

Vygotskii's and Leont'ev's reason for introducing the secondary stimulus (colored cards) into the paradigm was quite beneficial to advance performance; it served as a double stimulation technique which allowed children to rethink previous color assignments. The cards compelled children to consider not just what they were about to do, but to utilize double consciousness to reflect upon past steps when proceeding to subsequent ones. In short, children's use of an external stimulus as a physical tool (rather than depending upon adult instructions and/or memory alone) became a means to force children to depend upon conscious reflection – comparing and integrating all responses. Hence, Vygotskii's double stimulation device had the benefit of controlling children's future responses. It compelled children to mediate attention. In fact, (CW vol. 4: 155) only those children beyond 8;0 reliably utilized the colored cards as memory aids to preclude dependence upon automatic sign-object associations, employing substitute predicates. "Sometimes the child solves the problem completely differently. He does not put the forbidden colors aside, but selects them and puts them before him and fixes his eyes on them. In these cases, the external device corresponds precisely to the internal operation, and we have before us the operation of mediated attention" (Vygotskii CW vol. 4: 157). The decision-making at this stage is mediated by the extent of children's negative affect owing to expectations of how to control which outcome is more advantageous to them. Although some internal processes (in the form of conflict for the child alone), argumentational strategies are still not fully developed – children are motivated by ego-based outcomes.

In other words, before 8;0 subjects were not ordinarily able to solve a simple naming problem which required conflict of motive/double stimulation—they often were unable to substitute a less typical predicate (a description) for a more conventional one. Upon observing the same object a second time, subjects were unable to replace a more typical quality with a less typical one, e.g., "grass is "brown." It was not until

after 8;0 that children were successful at replacing typical with less typical attributes; their explanations included receipt of insufficient hydration to the grass (given the necessity of applying the color "brown" to the grass). Younger children's responses (between 4;0 and 6;0) revealed less success in making the secondary label substitution. The younger children's reticence to assign secondary linguistic signs and non-nucleus meanings to the same object/event may well be a consequence either of inability to utilize indexes effectively, or of dependence upon formulaic linguistic and cognitive associations. While the former inhibits semiosis by failure to apprehend morphemic differentiations when producing syntactic strings, the latter prevents new meaning assignments in resisting "decoupling" between already conceived interpretations of stimuli and the stimuli themselves (cf. Leslie 1987). Lack of means to supply rationale for exceptions to conceptual rules testifies to children's resistance – to recognize and associate different meanings to the same sign.

At 8;0 (and beyond) auxiliary stimuli become internal, requiring a process of increasing self-control to decide between courses of action – superseding the interventive effect of linguistic stimuli alone. Vygotskii intended this experimental paradigm to measure a third and higher level of double stimulation in which "conflict of motives" is emphasized. Measuring whether a conflict of motives exists entails use of a logical (mental) device which privileges argumentative structure by comparing distinctive consequences from different decisions/outcomes. In this way, the existence of conflicts of motives is equivocal to double stimulation and double consciousness, since it demonstrates that the same mind is reflecting upon diverse viewpoints. But these perspectives supersede different orientations to spatial arrays. Instead, their representations are logical in nature (in the form of arguments); and they drive children to use a different kind of auxiliary stimulus (less iconic) when selecting a course of action. Accordingly, to become effective instruments for points of view consideration, these tools must balance objective principles and their outcomes against individual (subjective) inclinations to arrive at a "better" course of action. These logical tools likewise need to gradually increase the number of alternative perspectives – to compel children to formulate more complex decisions, settling upon the most plausible approach. "The qualitative changes were evident in that the unequivocal motive was replaced by ambiguous motives and this resulted in a complex adjustment with respect to the given series of actions. …From the aspect of method, the substantial change [equivocal rather than ambiguous motives] introduced by this device [increasing possible alternatives for decision-making] consists in our being able to create motive experimentally since the series which we use are flexible and can be increased, decreased, replaced in part, and finally moved from series to series" (Vygotskii CW vol. 4 1931/1997: 207–208). In this third experimental design, Vygotskii is able to manipulate the number of conflicts to resolve in order to arrive at an ultimate decision for self or for the general other. This design demonstrates that older children still utilize and command the use of physical devices to enhance attention, memory, and conflict resolution; but, they depend upon more objective, mental artifacts to organize decisions: "Older school children use external devices most fully and most adequately; they no longer exhibit complete dependence on the cards [external stimuli] as do the younger children" (CW vol. 4: 156).

Instead, older children rely upon an "auxiliary stimulus," a neutral device which helps organize/regulate their decisions – how to change physical outcomes (Sannino 2015: 9–10; Vygotskii 1931/1997: 210). This third kind of experiment (Vygotskii's unique design), challenges children to "… recognize the need to make a choice based on motive and …his [the child's] freedom is the recognition of necessity. The child controls his selection reaction, but not in such a way as to change the laws that govern it," (Vygotskii CW vol. 4 1931/1997: 210). Children fail to modify the "rules" because they have not yet reached the internal objectivity necessary to impose logical regulation to the best course of action in view of the potential consequences. Sannino's (2015: 2) contention that the import of intentionality in the process of determining and planning successful strategies for self and for others is revisionary; it shows how double stimulation can be extended from intrasubjective problem-solving scenarios (utilizing physical prompts as memory tools to resolve conflicts) to intersubjective genres in which dialogic sign use is primary. As such, either the self presents alternative approaches to the self, or other players (via articulated dialogue) make semiotic meanings together. This process entails introducing to one another strategies (implicitly, explicitly) which have either been successful for the speaker, or which possess some potency for future success.

6 The Primacy of Language as Vygotskii's Semiotic Tool

Vygotskii's distinction between "signification" and "sense" further clarifies the properties lacking in private speech, and demonstrates how speech becomes internalized. For Vygotskii (1934/1986: 247–252), "signification" obviates meanings of single users housed within momentary external linguistic productions, while "sense" is equated with wider meanings, which are shared, and socially adopted. These wider meanings (sense) may be contained within dictionary entries (Vygotskii 1934/1986), and constitute meaning-changes within dialogic processes, such that shared meanings of words are more amplified; and the potential of diluting their individual meaning within discourse is a likely prospect. "A single word is so saturated with sense that, like the title *Dead Souls*, it becomes a concentrate of sense. To unfold it into overt speech, one would need a multitude of words (Vygotskii 1934/1986: 247). Here "sense" is equivocal to extended, objective meanings of words as they are contextualized with other words, consonant with Peirce's Logical Interpretant. The meaning is amplified to include that which is agreed upon by interlocutors of that culture, and hence reflects objective qualities. But, the fact that word meaning requires "a multitude of words" to express the "sense" emphasizes the need to utilize indexical components, specifically deictic determinants in the comprehension process. In short, as lexical entries, words possess a diverse character – greater than that of thought.

In fact, Vygotskii (1934/1986: 252) contends that thought is inferior to language, because it is incapable of growth from within itself, and is dependent upon another quite distinctive system to promote generativity: "Thought is not begotten by thought; it is engendered by motivation, i.e., by our desires and needs, our interests and emotions." Vygotskii takes the position that thought without emotion/motivation cannot give rise to thought. This line of reasoning indicates that affect is a necessary

component for thought generativity, given its direct, but invisible means to highlight opposing motivations/perspectives, which Vygotskii refers to as "conflict of motive" (Vygotskii CW vol. 4 1931/1997: 201–208); but, this perspective-taking facility is still inferior to the competencies needed to produce conventionally recognized words, and the sequential properties of syntax.

Thought's inferiority emanates from its vague character, in not establishing individual components of propositions and arguments (be they expressed overtly or covertly): "Thought, unlike speech, does not consist of separate units. When I wish to communicate the thought that today I saw a barefoot boy in a blue shirt running down the street, I do not see every item separately: the boy, the shirt, its blue color, his running, the absence of shoes. I conceive of all this in one thought, but I put it into separate words" (Vygotskii 1934/1986: 251). In thought, elements of propositions may remain unnoticed, because their subjects and predicates are not individuated; hence, the precision of terms within the proposition are unrecognized, without the word. For example, in not recognizing the color (blue) or the physical state (barefoot) of the shirt/boy, the proposition is altered; and the producer's intent fails to be communicated.

Because of the imprecision of thoughts, in failing to make plain qualitative determinates of terms, they are subject to misinterpretation. This is especially so between interlocutors, given that the producer's meanings are not direct (Vygotskii 1934/1986: 252) – in that they are not expressed by a tangible sign. As a consequence of the intangible nature of the thought-sign, meanings are often inaccurately associated with terms which were unintended. Accordingly, it is the word (through its individuating (or limiting) function) which forces interlocutors to attend to and to become conscious of each element of the proposition, tightening the propositions themselves. Hence, "Thought must first pass through meanings and only then through words" (Vygotskii 1934/1986: 252). But, the clarification of meanings follows a discrete process: implementation of overt signs, together with meaning limitations. To orchestrate this transition (from thought to word), "word meanings must be cut" (Vygotskii 1934/1986: 251). The meanings which are "cut" entail members which were not intended to be included in the term of the proposition, e.g., shoe-clad boys or red shirts. As such, transitioning to the word forces interpreters, by overt symbols, to take notice of to whom or about what the proposition actually pertains. Vygotkskii describes this process of thought becoming as: "realizing itself in words" (Vygotskii 1934/1986: 251). This realization requires that an unconscious memory is attended to in such a way that it becomes a conscious proposition/assertion.

Vygotskii provides further rationale for the consciousness-raising power of the word:

> If perceptive consciousness and intellectual consciousness reflect reality differently, then we have two different forms of consciousness. *Thought and speech turn out to be the key to the nature of human consciousness.* If language is as old as consciousness itself, and if language is a practical consciousness-for-others and, consequently, consciousness-for-myself, then not only one particular thought but all consciousness is connected with the development of the word. The word is a thing in our consciousness,… that is absolutely impossible for one person, but that becomes a reality for two. The word is a direct expression of the historical nature of human consciousness. Consciousness is reflected in a word as the sun in a drop of water. A word relates to consciousness as a living cell relates to a whole organism, as an atom relates to the universe. A word is a microcosm of human consciousness. (Vygotskii 1934/1986: 256)

This materializes from the fact that words imply logical relations via propositions/assertions/arguments. "The thought is not only externally mediated by signs [words], but also internally by meanings. The whole point is that direct communication of minds is impossible not only physically, but also psychologically. It can only be reached through indirect, mediated ways. This road amounts to the internal mediation of the thought first by meanings, then by words. Therefore, the thought can never be equal to the direct meaning of words. The meaning mediates the thoughts on its road towards verbal expression, that is, the road from thought to the word is a roundabout, internally mediated road" (Vygotskii 1934/1986: 314). van der Veer (1991: 369) comments that between thought and word "sense dominates over meaning;" and this more objective logical state is characterized as inner speech. In this way, inner speech informs reasoning by practicing logical paradigms; and as such, inner speech improves all kinds of reasoning: abductive, inductive, and deductive. The self discourses with the self the soundness of newly-conceived hypotheses, thereby agent and patient discourse within a single individual.

Nonetheless, Vygotskii emphasizes that it is the element of thought that is driven by emotion and conflicts of motive that provides its objective character; hence when it acquires this objective element, it is afforded a primary function – although this function still lacks the means to self-regulate. Vygotskii's notion of thought constitutes an event repository from which the self can organize and then determine which logical relations have merit. Examination of conflicts of motive, which regulates conduct, requires cultural tools beyond mere thought, namely, inner speech, and other attention inducers such as temporal and spatial memory aids (watches, die, color displays). Conflicts of motive form the catalyst for relating contributory events to potential consequences. Beyond the use of physical memory aids, words memorialize, then regulate responses. It is the "hidden meanings of the words" which illustrate the impact of thought on language and behavior, because each spoken line has an underlying "motive" (van der Veer and Valsiner 1991: 370). It is obvious then that the necessary component of conflict of motive (which incorporates affect) demonstrates that achieving inner speech is a process which relies heavily upon meaning alterations, and hence incorporates a key component of sign use, meaning/effect. For this reason, the nature of Vygotskii's conflict of motive prominently reveals a semiotic process (cf. van der Veer 1991: 370 for further discussion). For Vygotskii (1934/1986: 269), "thought is not expressed in the word, but is completed in the word." van der Veer and Valsiner (1991: 370) express this as: "…becoming (the unity of being and non-being) of the thought in the word." Although thought motivates speech, its substance and perhaps its structure (syntax and logical sequencing) are substantially changed by producing words (cf. van der Veer and Valsiner 1991: 371 for a foundational account).

7 Peirce's Double Consciousness

By double consciousness, Peirce refers to a form of dialogue (either within the same individual or between different individuals) whose effect results in increased awareness of diverse perspectives pertaining to the truth of propositions/arguments. Its characteristics are consonant with narrative structure/meaning, because both rely upon reflection of episodic sequences which are elementary components of propositions/arguments

(cf. Bamberg 1992, 1997; West under review). The onset of narratives can measure the point in development when double consciousness emerges – when children acquire sufficient reflective skills to appreciate mutual perspectives, and begin to exploit such through overt or covert linguistic exchanges.

Not until children's narration is characterized as folk psychological, can they truly narrate events which draw upon what Peirce refers to as "common ground" (1906: MS 614) – bringing to bear the dialogical element of double consciousness. Doing so entails appreciating distinctions in viewpoints, as well as the explanations which underlie them (Gallagher and Hutto 2012: 30). The emergence of folk psychological narratives is compelled by "…the actions of others [when they] deviate from what is normally expected in such a way that we encounter difficulties understanding them" (Gallagher and Hutto 2012: 30). Allocentric viewpoints are apprehended especially when they surface unexpectedly and when their character is distinctive in terms of another's conduct, reminiscent of Peirce's concept of double consciousness and Vygotskii's notion of double stimulation "a conflict of motive" (cf. Sannino 2015: 9 and Vygotskii 1931/1997: 152–158). The conflict arising from another's unexpected conduct (linguistic, non-linguistic) makes it "notable," causing the conduct to "fall into the spotlight for special attention and explanation" (Gallagher and Hutto 2012: 30). In fact, reflecting upon another's seemingly anomalous behavior impels consideration of the rationale underlying such behavior, compelling "…explanations of a specific sort that involve understanding the other's reasons for taking the particular action" (Gallagher and Hutto 2012: 30). The rationale is as follows: others' unexpected actions/beliefs beget conflict, which in turn, begets a certain graduated level of conscious awareness, rising to the level of reflective/metacognitive considerations. This form of awareness supersedes unconscious awareness (noticing the absence of something), since the element of surprise upon encountering the conflict is sufficiently notable that attention to its differentness necessarily materializes. These more conscious, reflective skills advance children's means to recognize viewpoints beyond their repertoire (given their differences), and to change original perspectives if appropriate. Changes in originary perspectives initially take the form of viable hunches, such that proposals surface to explain anomalous circumstances.

According to Peirce's concept of double consciousness, this reflective competence supplies the nuts and bolts to master inner dialogue – self-talk in which conflicts are resolved, and explanatory hypotheses are reconciled between two phases of the self: "…thinking always proceeds in the form of a dialogue – a dialogue between phases of the ego – so that, being dialogical, it is essentially composed of signs as its matter, in the sense in which a game of chess has the chessmen for its matter" (1906: 4.6). Peirce's analogy of the game of chess in which "chessmen" are "its matter" describes how dialogue is, by nature, a semiotic system which could not materialize but for the association of common meanings with the same representations. Both processes (playing chess and dialogue) require planning; and planning requires a strategy of predictions derived from expectations of another's responses. This is precisely what Peirce means by "common ground/place to stand." Meaning extensions are transferred either from one party to another; or new meanings are communicated from one phase of the ego to another phase (provided that enough shared foundational sign-object meanings are in place).

Peirce's double consciousness incorporates this element of common knowledge and common expectations. In this way it bears the primary characteristics of inner dialogue ("phases of the ego"), consonant with Vygotskii's "inner speech." But what is distinctive in Peirce's account of inner dialogue (as previously alluded to) is the component of surprise upon experiencing a percept, and strong resistance between two contrastive percepts/propositions/arguments (cf. 1903: 5.53). The component of surprise implicates the presence of abductive reasoning, in that resolving the surprise/unexpected event with previous related assumptions requires the generation of workable hunches to explain how the unexpected can plausibly surface, and make sense in the context. In fact, even conflicting percepts can promote surprise, and can serve as the raw material for the give and take of arguments inherent in double consciousness. This surfaces when percepts contain interpretations; and the presence of interpretations demonstrates that implied propositions/arguments are housed within the percept as Semes (1906: 4.538, MS 295: 26).[3] The Seme arises from the "widening" (4.538) of Term, Proposition, and Argument, in order to make the first two a division of all signs.

In fact, the element of resistance intrinsic to double consciousness has its foundation in hunches which begin as unexpected images/percepts which already qualify as signs, given their association with meaning. This phenomenon epitomizes the resistance so inherent in the emergence of moments of insight flowing from unexpected happenings – they materialize particularly upon the sudden flash of a vivid, but dissonant, percept. Peirce's emphasis on vividness demonstrates the primacy of the phenomenological within dialogic reasoning, because it entails idiosyncratic affect – what is vivid to one observer may not be so for another. Such is not merely an empirical construct. The fact that "vividness is exalted" (1903: 5.53) further affords percepts their phenomenological status. The affect which draws attention to percepts constitutes but one side of double consciousness, warring with other established affect-based affinities. Accordingly, vivid percepts represent the pivotal component to obviate the two critical sides of double consciousness:

> "Examine the Percept in the particularly marked case in which it comes as a surprise. At the moment when it was expected the vividness of the representation is exalted…something quite different comes instead. I ask you whether at that instant of surprise that there is not a double consciousness, on the one hand of an ego, which is simply the expected idea suddenly broken off, on the other hand of the non-ego, which the strange intruder, in his abrupt entrance." (1903: 5.53; likewise cf. 1908: SS 195)

The fact that the "strange intruder" can surface as a vivid percept is a prime illustration of Peirce's characterization of the "Firstness of Secondness" (cf. Atkins 2018). Peirce indicates that the "abrupt entrance" of the "strange intruder" illustrates an insistency and persistency of the new idea (cf. also 1906: MS 298: 29–31); and the category of the Firstness of Secondness has a marked influence in establishing double

[3] "By a *Seme*, I shall mean anything which serves for any purpose as a substitute for an object of which it is, in some sense, a representative or Sign. The logical Term, which is a class-name, is a Seme" (4.538).

consciousness, given the impingement of the objects' qualities upon the interpreter (cf. Atkins 2018: 197–198).[4]

Peirce refers to this inpingement of the object's qualities (Firstness of Secondness) upon the interpreter as "perceptuation": "A perceptuation is an externisensation in which the active element is volitionally external while the passive element is volitionally internal" (1905: MS 339: 245r). It is obvious that Peirce gives great weight to the object, its qualities, together with its means to compel attention – obviated when he characterizes the perceptuation process (double consciousness) as more "involuntary" than "voluntary" (1903: 331; 1905: MS 339: 245R). The "voluntary" element represents factors internal to ego; whereas "involuntary" influences pertain to external forces, such as objects. Hence, the process of double consciousness and its effects depend more upon external than internal influences. Perceptuation is involuntary, in that the object and its properties are responsible for noticing, and becoming aware of the object and for interpreting its meaning/influence upon states of affairs. Firstness enters in when the object's intensity accounts for the feelings which surface upon observing it: Peirce characterizes this feeling of Firstness within the mantle of double consciousness, because the feeling is double-sided (representing diverse perspectives) and emphasizes the newness of feeling, together with its possible consequences (Atkins 2018: 212–213). Atkins (2018: 214) further indicates that Peirce's double consciousness extends into the realm of the imagination, affecting the course of meanings in all possible instantiations and in all contexts to come. In view of these applications, Peirce demonstrates his commitment to the future effects of double-sided feeling from the object's insistence in Secondness as being: "produced by the reaction of the ego on the non-ego" (1906: MS 298: 296). Hence, double consciousness requires a change by virtue of a brute conflict between old and new perspectives – a sudden encounter between what the self already knows, and the novel interpretation of the not-self, which insinuates itself ("as an intruder") upon the consciousness (1903: 5.53).

Atkins (2018) analogizes objects' effects upon the consciousness with those of ideas. He elaborates how objects can make themselves known in a quantitative manner, when they perpetuate notice of the intensity present in their quality (of the impression). "The total intensity of a color is a function of its relative intensities, so too is the total clarity of an idea a function of the relative degrees of attainment with respect to clearness, distinctness, and pragmatic adequacy" (Atkins 2018: 202) This impression of intensity reveals the objects' own qualities, while vividness arises out of the perceptual experience of the observer (cf Atkins 2018: 212 for further discussion) – obviating the interpretive effects upon the perceiver, which evolve with time and experience. Furthermore, this feature of vividness can surface in both visual and non-visual modalities, as Atkins (2018: 217–218 makes plane), e.g., a blaring trumpet, or the red color of a fire engine. The vividness of the qualities of these objects (in some cases affected by the object's intensity) are so riveting that the "mental eyeballs" of the observer are forced upon the objects' (1908: 8.350), such that the intensity of the quality gives rise to

[4] "An externisensation is a two-sided state of determination of consciousness…in which a volitionally external and a volitionally internal power seem to be opposed, the one furthering the other resisting the change" (MS 339: 245r).

"secondary feelings (1906: MS298: 296). Atkins (2018: 223) equates this with synesthesia, the idea that competing modalities can provoke similarly intense effects, or "vividness," as Peirce terms it in 1905 (MS 298: 29).

In showcasing transcendence from surprise and sudden insight in the percept to more deliberately found and more analyzed perceptual judgements, Peirce further illustrates the relevance of Secondness – particularly as resistance to novel perspectives contained within the double-sided paradigm of double consciousness. While the percept ordinarily surfaces as a sudden image – with sensations flashing across the consciousness, afterward it becomes a more analyzable mental reaction in the perceptual judgement (cf. Short 2007: 319 and Wilson 2016: 96–97). In other words, a mental image becomes a perceptual judgement upon its interpretation, triggered by resistance in the face of surprise: "Examine the percept in the particularly marked case in which it comes to us as a surprise. At the moment when it was expected the vividness of the representation is exalted…" (1903: 5.53). The representation's exaltation (its vividness) results from the element of surprise – surfacing with the realization that relations exist between two or more events. The element of surprise materializes upon notice that a novel schema is to be applied to explain the event relations implicated by the seemingly anomalous percept; and Attention to the feeling-based consequence further heightens the new, intruding percept, especially given its unanticipated but fitting relevance. As a consequence, the place of the percept within an episode acquires status as a perceptual judgement, given its interpretation as a logical link within an event structure (cf. West 2014). It is at this juncture that the percept is inaugurated as a sign, given the assignment of meaning. The percept becomes a sign upon its vivid notice, when it is connected with an interpretation; and the element of surprise demarcates and illustrates the meanings which intrude, suddenly infusing the sign.

In the same passage, Peirce follows with a description of the reaction which comprises the surprise: "Something quite different comes instead. I ask you whether at that instant of surprise that there is not a double consciousness, on the one hand of an ego, which is simply the expected idea suddenly broken off, on the other hand the non-ego, which is the strange intruder, in his abrupt entrance" (1903: 5.53). The something that is expected is replaced by a quite different something. Peirce's concept of double consciousness characterizes not merely the initial inconsonance/clash between the old and new meanings, but the new belief/action habit implemented consequent to the clash. The new belief/action pattern constitutes the "strange intruder;" it "intrudes" when novel interpretations/meanings intervene, such that the new hunch takes the place, obliterating the previously established cognition/action pattern.

To further highlight the effect of double consciousness, Peirce institutes a taxonomy emphasizing the inauguration of meaning in the process of perceiving percepts, calling it "percipuum." Double consciousness is established by the process of associating meanings with vivid images – constituting the subject, then, by means of the perceptual judgment, generating a proposition with the application of a predicate. Gradually, the percipuum individuates the point when the image is accorded definitive interpretation, and is about to qualify as a perceptual judgement: "But at that moment, [of surprise] that one fixes one's mind upon it [the new meaning of a percept]" (1903: 7.643), is the very moment when one attributes to the percept some new meaning/habit-change. The percept

becomes a "percipuum" en route to a perceptual judgement, when an anomalous relation between it and its effect infuses the percept with amplified significative power:

> "…it is in conflict with the facts which are that a man is more or less placidly expecting one result, and suddenly finds something in contrast to that forcing itself upon his recognition. A duality is thus forced upon him: on the one hand, his expectation which he had been attributing to Nature, but which he is now compelled to attribute to some more inner world, and on the other hand, a strong new phenomenon which shoves that expectation into the background and occupies its place. The old expectation, which is what he was familiar with, is his inner world, or ego. The new phenomenon, the stranger, is from the exterior world or non-ego." (1903: EP2:195)

Recognition of the value of "the stranger" to mutual consideration of novel propositions/arguments constitutes the impetus for double consciousness.

8 Conclusion

Peirce's double consciousness constitutes an overarching tool to promote inferential reasoning – amplifying Vygotskian-based interventions of: private speech, collaborative speech (between interlocutors), and inner speech. Because sign exchange underlies Peirce's model of metaphysics and phenomenology, it highlights how indexical and iconic signs (signs not constrained to explicit, symbolic signs) facilitate the awareness of perspectives beyond the subjective, e.g., reciprocal shifting of attention to alternative focus. Peirce's double consciousness constitutes the logico-linguistic tool to foster just such dialogic advances. It orchestrates this "…because of…material constraints that the internal mind, so to speak, can 'learn' and unshackle its own rigidities and impossibilities…" (Magnani 2018: 150). Moreover, Peirce privileges the element of surprise within dialogic exchanges (linguistic, and non-linguistic alike) by means of the imposition of the "strange intruding" idea. His notion of double consciousness obviates perspective shifts by way of sign use which requires and encourages uncovering implicit meanings. As Peirce indicates, even percepts have the potential to imply propositions/arguments, provided that interpretation and habit-change are fertile. The value of interpretive percepts as forums to lay bare contrasting perspective within venues of double consciousness, is to allow for the give-and-take of legitimate hunches, prior to private and inner speech. As such, vivid mental pictures can serve as early tools to compel inferential reasoning, even in pragmatic pursuits. In not privileging linguistic signs to advance dialogic exchanges, Peirce uncovers a wealth of enhancements to problem-solving which supersedes the influence of language advocated by Vygotskii.

The episodic features which Peirce's index and icon afford in representing novel hunches to the self, prior to the onset of language, together with its means to augment language meanings is indispensable in providing early practice in developing and settling upon workable propositions/arguments. Peirce's maxim – that the universe must be "well known and mutually known to be known" (1901: 3.621) is facilitated by percepts in the form of virtual habits (vivid pictorial images), and acts of nature as commands (MS 295), provided that the same implied propositions/arguments are assumed between minds or actualized within ego's mind. In this way, Peirce's model

affords percepts along with conduct, and words, a role in dialogic exchanges, as long as their meaning contributes some renovative propositional/argumentative promise. In short, double consciousness can intervene in the realm of images or in linguistic genres, to (via the addition of predicates) impose propositions/arguments (implicit or explicit). Hence, double consciousness forums can materialize either intrasubjectively (self talking to self audibly or silently); or it can emerge as an intersubjective phenomenon – two distinctive minds sharing propositions and arguments after vivid percepts become interpreted in percipua or in perceptual judgements.

References

Acredolo L, Goodwyn S (1990) Sign language in babies: the significance of symbolic gesturing for understanding language development. In: Vasta R (ed) Annals of child development, vol 7. JAI Press, Greenwich, pp 1–42

Alderson-Day B, Fernyhough C (2015) Inner speech: development, cognitive functions, phenomenology, and neurobiology. Psychol Bull 141(5):931–965

Atkins RK (2018) Charles S. Peirce's phenomenology: analysis and consciousness. Oxford University Press, Oxford

Austin JL (1962) How to do things with words. In: Urmson JO, Sbisà M (eds): The William James Lectures delivered at Harvard University in 1955. Clarendon Press, Oxford

Baldwin DA, Saylor M (2005) Language promotes structured alignment in the acquisition of mentalistic concepts. In: Wilde J, Baird JA (eds) Why language matters for theory of mind. Oxford University Press, Oxford, pp 123–143

Bamberg M (1997) A constructivist approach to narrative development. In: Bamberg M (ed) Narrative development: six approaches. Lawrence Erlbaum Associates, Hilldale, pp 89–132

Bamberg M (1992) Binding and unfolding. establishing viewpoint in oral and written discourse. In: Kohrt M, Wrobel A (eds) Schreibprozesse - Schreibprodukte: Festschrift für Gisbert Keseling. Georg Olms Verlag, Hildesheim

Bellucci F (2014) Logic, considered as semeiotic: on Peirce's philosophy of logic. Trans Charles S. Peirce Soc 50(4):523–547

Bergman M (2016) Habit-change as ultimate interpretant. In: West D, Anderson M (eds) Consensus on Peirce's concept of habit: before and beyond consciousness. Springer, Heidelberg

Cameron CA, Lee K (1997) The development of children's telephone communication. J Appl Dev Psychol 18:55–70

Cameron CA, Wang M (1999) Frog, where are you? Children's narrative expressions over the telephone. Discourse Process 28(3):217–236

Cameron CA, Hutchison J (2009) Telephone-mediated communication effects on young children's oral and written narratives. First Lang 29(4):347–371

Clark E (2009) First language acquisition. Cambridge University Press, Cambridge

Deely J (2009) Purely objective reality. Walter de Gruyter, Berlin

Deely J (2012) Toward a postmodern recovery of "person". Espiritu 61(143):147–165

Fernyhough C (2008) Getting Vygotskian about theory of mind: mediation, dialogue, and the development of social understanding. Dev Rev 28:225–262

Fivush R, Haden C (1997) Narrating and representing experience: preschoolers' developing autobiographical accounts. In: van dan Broek P, Bauer P, Bourg T (eds) Developmental spans in event comprehension and representation. Lawrence Erlbaum Associates, Hilldale, pp 169–198

Gallagher S (2017) Enactivist interventions: rethinking the mind. Oxford University Press, Oxford

Gallagher S, Hutto D (2012) Understanding others through primary interaction and narrative practice. In: Zlatev J, Racine T, Sinha C, Itkonen E (eds) The shared mind: perspectives on intersubjectivity. John Benjamins, Amsterdam, pp 17–38

Goudena P (1992) The problem of abbreviation in internalization of private speech. In: Diaz R, Berks L (eds) Private speech: from social interaction to self-regulation. Lawrence Erlbaum Associates, Hillsdale, pp 215–224

Halliday MAK, Hasan R (1976) Cohesion in English. Longman Group Ltd, London

Hayne H, Imuta K (2011) Episodic memory in 3 and 4-year-old children. Dev Psychol 53:317–322

Hyams N (1994) V2, null arguments and COMP projections. In: Hoekstra T, Schwartz B (eds) Language acquisition studies in generative grammar. John Benjamins, Amsterdam, pp 21–56

Kilpinen E (2016) In what sense exactly is Peirce's habit-concept revolutionary? In: West D, Anderson M (eds) Consensus on Peirce's concept of habit: before and beyond consciousness. Springer, Heidelberg, pp 199–213

Leslie AM (1987) Pretense and representation: the origins of "theory of mind". Psychol Rev 94 (4):412–426

Lust B, Chien Y-C, Flynn S (1987) What children know: methods for the study of first language acquisition. In: Lust B (Ed) Studies in the acquisition of anaphora, applying the constraints, vol 2. Reidel, Dordrecht, pp 272–356

McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago

Magnani L (2017) The abductive structure of scientific creativity: an essay on the ecology of cognition. Springer, Heidelberg

Magnani L (2018) Ritual artifacts as symbolic habits. Open Inf Sci 2018(2):147–155

Mayer M (1969) Frog, where are you?. Dial Press, New York

Montrul S (2004) The acquisition of Spanish: morphosyntactic development in monolingual and bilingual L1 acquisition and adult L2 acquisition. John Benjamins, Amsterdam

Neisser U (2004) Memory development: new questions and old. Dev Rev 24:154–158

Peirce CS (i. 1867–1913) Collected papers of Charles Sanders Peirce. In: Hartshorne C, Weiss P (ed) Harvard University Press, Cambridge, vol 1–6. In: Burks A (ed) Harvard University Press, Cambridge, vol 7–8. (1931–1966)

Peirce CS (i. 1867–1913) The essential Peirce: selected philosophical writings, vol 1. In: Houser N, Kloesel C (eds) vol 2 Peirce edition project. University of Indiana Press, Bloomington. (1992–1998)

Peirce CS (i. 1867–1913) Unpublished manuscripts are dated according to the annotated catalogue of the papers of Charles S. Peirce. In: Robin R (ed). University of Massachusetts Press, Amherst. Confirmed by the Peirce Edition Project (Indiana University-Purdue University at Indianapolis). (1967)

Peirce CS, Welby V (i. 1898–1912) Semiotic and significs: the correspondence between Charles S. Peirce and Victoria, Lady Welby. In: Hardwick C, Cook J (eds) University of Indiana Press, Bloomington. (1977)

Perner J, Ruffman T (1995) Episodic memory and autonoetic consciousness: developmental evidence and a theory of childhood amnesia. J Exp Child Psychol 59:516–548

Pietarinen A-V (2006) Signs of logic: Peircean themes of the philosophy of language, games, and communication. Springer, Heidelberg

Pillemer DB, White S (1989) Childhood events recalled by children and adults. Adv Child Dev Behav 21:297–340

Pinto G, Tarchi C, Gamannossi B, Bigozzi L (2016) Mental state talk in children's face-to-face and telephone narratives. J Appl Dev Psychol 44:21–27

Pinto G, Tarchi C, Bigozzi L (2018) Is two better than one? Comparing children's narrative competence in an individual versus joint storytelling task. Soc Psychol Educ 21:91–109
Rowlands M (2006) Body language. MIT Press, Cambridge
Sannino A (2015) The principle of double stimulation: a path to volitional action. Learn Culture Soc Interact 6:1–15
Saylor M (2004) Twelve- and 16-month-old infants recognize properties of mentioned absent things. Dev Sci 7(5):599–611
Short TL (2007) Peirce's theory of signs. Cambridge University Press, Cambridge
Stanfield C, Williamson R, Özçalişkan S (2014) How early do children understand gesture-speech combinations with iconic gestures? J Child Lang 41:462–471
Stjernfelt F (2014) Natural propositions: the actuality of Peirce's doctrine of dicisigns. Docent Press, Boston
Trabasso T, Stein N (1997). Narrating, representing, and remembering event sequences. In: van dan Broek P, Bauer P, Bourg T (eds.) Developmental spans in event comprehension and representation. Lawrence Erlbaum Associates, Hillsdale, pp 237–270
Tulving E (2005) Episodic memory and autonoesis: uniquely human? In: Terrace HS, Metcalfe J (eds) The missing link in cognition: origins of self-reflective consciousness. Oxford University Press, Oxford, pp 3–56
Valian V (1991) Syntactic subjects in the early speech of American and Italian children. Cognition 40:21–81
Valian V, Aubry S (2005) When opportunity knocks twice: two-year-olds' repetition of sentence subjects. J Child Lang 32:617–641
van der Veer R, Valsiner J (1991) Understanding Vygotsky: a quest for synthesis. Blackwell, Oxford
Vygotskii LS (i. 1924–1934) Thinking and concept formation in adolescence. In: van der Veer R, Valsiner J (eds) The Vygotsky reader. Blackwell, Oxford, pp 185–265. (1994)
Vygotskii LS (1931/1997) The history of the development of higher mental functions (collected works of vygotsky, vol 4). In: Rieber R (ed). Springer, Heidelberg
Vygotskii LS (1934/1986). Thought and language. (Trans: Kozulin A). MIT Press, Cambridge
West D (2011) Deictic use as a threshold for imaginative thinking: a peircean perspective. Soc Semiot 21(5):665–682
West D (2013) Deictic imaginings: semiosis at work and at play. Springer, Heidelberg
West D (2014) Perspective switching as event affordance: the ontogeny of abductive reasoning. Cogn Semiot 7(2):149–175
West D (2015) Dialogue as habit-taking in Peirce's continuum: the call to absolute chance. Dialogue (Can Rev Philos) 54(4):685–702
West D (2016) Indexical scaffolds to habit-formation. In: West D, Anderson M (eds) Consensus on Peirce's concept of habit. Springer, Heidelberg
West D (2017) Virtual habit as episode-builder in the inferencing process. Cognit Semiot 10 (1):55–75
West D (2018) Early enactments as submissions toward self-control: Peirce's ten-fold division of signs. In: Owens G, Pelkey J (eds) Semiotics 2017. https://doi.org/10.5840/cpsem20171
West D (Under review) Index as scaffold to the subjunctivity of children's performatives. Am J Semiot
West D (2019) Index as scaffold to logical and final interpretants: compulsive urges and modal submissions. Semiotica 228:333–353
Wilson A (2016) Peirce's empiricism: its roots and its originality. Lexington Books, Lanham
Winsler A, Diaz R, Montero I (1997) The role of private speech in the transition from collaborative to independent task performance in young children. Early Child Res Q 12:59–79



Creative Model-Based Diagrammatic Cognition

The Discovery of the "Imaginary" Non-Euclidean Geometry

Lorenzo Magnani(✉)

Department of Humanities, Philosophy Section, and Computational Philosophy Laboratory, University of Pavia, Pavia, Italy
lmagnani@unipv.it

Abstract. The present article is devoted to illustrate the issue of the model-based and extra-theoretical dimension of cognition from the perspective of the famous discovery of non-Euclidean geometries. This case study is particularly appropriate because it shows relevant aspects – creative – of diagrammatic cognition, which involve intertwined processes of both explanatory and non-explanatory abduction. These processes act at the model-based level taking advantage of what I call *mirror* and *unveiling* diagrams. A description of important abductive heuristics is also provided: expansion of scope strategy, Euclidean/non-Euclidean model matching strategy, consistency-searching strategy.

Keywords: Non-Euclidean geometry · Model-based reasoning · Diagrammatic reasoning · Geometrical construction · Manipulative abduction · Mirror diagrams · Unveiling diagrams · Mental models · Internal and external representations · Expansion of scope strategy · Euclidean/non-Euclidean model matching strategy · Consistency-searching strategy

1 Geometrical Construction Is a Kind of Manipulative Abduction

A traditional and important example of model-based reasoning is represented by the cognitive exploitation of diagrams. This article will deal with the classical case of diagrammatic reasoning in geometrical discovery, taking advantage of the discovery of the first non-Euclidean geometry, also called imaginary geometry (or hyperbolic geometry), due to N. J. Lovachevsky. Let's quote an interesting passage by Peirce about constructions. Peirce says that mathematical and geometrical reasoning "[...] consists in constructing a diagram according to a general precept, in observing certain relations between parts of that diagram not explicitly required by the precept, showing that these relations will hold for all such diagrams, and in formulating this conclusion in general terms. All valid necessary reasoning is in fact thus diagrammatic" (Peirce 1958, 1.54). Not

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 217–244, 2019.
https://doi.org/10.1007/978-3-030-32722-4_13

dissimilarly Kant says that in geometrical construction "[. . .] I must not restrict my attention to what I am actually thinking in my concept of a triangle (this is nothing more than the mere definition); I must pass beyond it to properties which are not contained in this concept, but yet belong to it" (Kant 1929, A718-B746, p. 580).

Manipulative abduction[1] is a kind of, usually model-based, abduction that exploits external models endowed with delegated (and often implicit) cognitive roles and attributes.[2] 1. The model (diagram) is external and the strategy that organizes the manipulations is unknown a priori. 2. The result achieved is new (if we, for instance, refer to the constructions of the first creators of geometry), and adds properties not contained before in the concept (the Kantian to "pass beyond" or "advance beyond" the given concept (Kant 1929, A154-B194, p. 192)).[3]

Humans and other animals make a great use of perceptual reasoning and kinesthetic and motor abilities. We can catch a thrown ball, cross a busy street, read a musical score, go through a passage by imaging if we can contort out bodies to the way required, evaluate shape by touch, recognize that an obscurely seen face belongs to a friend of ours, etc. Usually the "computations" required to achieve these tasks are not accessible to a conscious description. Mathematical reasoning uses language explanations, but also non-linguistic notational devices and models. Geometrical constructions represent an example of this kind of extra-linguistic machinery we know as characterized in a model-based and manipulative – abductive – way. Certainly a considerable part of the complicated environment of a thinking agent is internal, and consists of the proper software composed of the knowledge base and of the inferential expertise of that individual. A cognitive system consists of a "distributed cognition" among people and "external" technical artifacts (Hutchins 1995; Zhang 1997).

In the case of the construction and examination of diagrams in geometry, a sort of specific "experiments" serve as states and the implied operators are the manipulations and observations that transform one state into another. The mathematical outcome is dependent upon practices and specific sensorimotor activities performed on a non-symbolic object, which acts as a dedicated external

[1] Abduction refers to all kinds of reasoning to hypotheses, especially explanatory, as Charles Sanders Peirce illustrated.

[2] The concept of *manipulative abduction* – which also takes into account the external dimension of abductive reasoning in an eco-cognitive perspective – captures a large part of common and scientific thinking where the role of action and of external models (for example diagrams) and devices is central, and where the features of this action are implicit and hard to be elicited. Action can provide otherwise unavailable information that enables the agent to solve problems by starting and by performing a suitable abductive process of generation and/or selection of hypotheses. Manipulative abduction happens when we are thinking through doing and not only, in a pragmatic sense, about doing (cf. (Magnani 2009, chapter one)).

[3] Of course in the case we are using diagrams to demonstrate already known theorems (for instance in didactic settings), the strategy of manipulations is already available and the result is not new.

representational medium supporting the various operators at work. There is a kind of an epistemic negotiation between the sensory framework of the mathematician and the external reality of the diagram. This process involves an external representation consisting of written symbols and figures that are manipulated "by hand". The cognitive system is not merely the mind-brain of the person performing the mathematical task, but the system consisting of the whole body (cognition is *embodied*) of the person plus the external physical representation. For example, in geometrical discovery the whole activity of cognition is located in the system consisting of a human together with diagrams.

An external representation can modify the kind of computation that a human agent uses to reason about a problem: the Roman numeration system eliminates, by means of the external signs, some of the hardest parts of the addition, whereas the Arabic system does the same in the case of the difficult computations in multiplication (Zhang 1997). All external representations, if not too complex, can be transformed in internal representations by memorization. But this is not always necessary if the external representations are easily available. Internal representations can be transformed in external representations by externalization, that can be productive "[. . .] if the benefit of using external representations can offset the cost associated with the externalization process" (Zhang 1997, p. 181). Hence, contrarily to the old view in cognitive science, not all cognitive processes happen in an internal model of the external environment. The information present in the external world can be directly picked out without the mediation of memory, deliberation, etc. Moreover, various different external devices can determine different internal ways of reasoning and cognitively solve the problems, as is well-known. Even a simple arithmetic task can completely change in presence of an external tool and representation. In the Fig. 1 an ancient external tool for division is represented.

Following the approach in cognitive science related to the studies in distributed cognition, I contend that in the construction of mathematical concepts many external representations are exploited, both in terms of diagrams and of symbols. I have been interested in my research in diagrams which play an *optical* role[4]– microscopes (that look at the infinitesimally small details), telescopes (that look at infinity), windows (that look at a particular situation), a *mirror* role (to externalize rough mental models), and an *unveiling* role (to help create new and interesting mathematical concepts, theories, and structures).

Moreover optical diagrams play a fundamental explanatory (and didactic) role in removing obstacles and obscurities (for example the ambiguities of the concept of infinitesimal)[5] and in enhancing mathematical knowledge of critical situations (for example the problem of parallel lines, cf. the following sections). They facilitate new internal representations and new symbolic-propositional achievements. The mirror and unveiling diagrammatic representation of mathematical structures activates *perceptual operations* (for example identifying the

[4] This method of visualization was invented by Stroyan (2005) and improved by Tall (2001).

[5] Cf. chapter one, section 1.7 of my book (Magnani 2009).

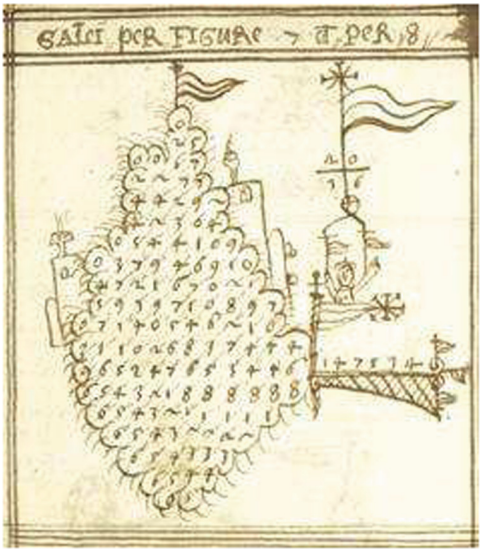


Fig. 1. Galley division, XVI Century, from an unpublished manuscript of a Venetian monk. The title of the work is Opus Artimetica D. Honorati veneti monachj coenobij S. Lauretij.

interplay between conflicting structures: for example how the parallel lines behave to infinity). These perceptual operations provide mirror and unveiling diagrammatic representations of mathematical structures.

To summarize we can say mathematics diagrams play various roles in a typical abductive way; moreover, they are external representations which, in the cases I will present in the following sections, are devoted to provide explanatory and non-explanatory abductive results. Two of them are central:

- they provide an intuitive and mathematical *explanation* able to help the understanding of concepts difficult to grasp or that appear obscure and/or epistemologically unjustified. I will present in the following section some mirror diagrams which provided new mental representations of the concept of parallel lines.
- they help abductively *create* new previously unknown concepts that are *non-explanatory*, as illustrated in the case of the discovery of the non-Euclidean geometry.

2 Mirror Diagrams: Externalizing Mental Models to Represent Imaginary Entities

Empirical anomalies result from data that cannot currently be fully explained by a theory. They often derive from predictions that fail, which implies some

element of incorrectness in the theory. In general terms, many theoretical constituents may be involved in accounting for a given domain item (anomaly) and hence they are potential points for modification. The detection of these points involves defining which theoretical constituents are employed in the explanation of the anomaly. Thus, the problem is to investigate all the relationships in the explanatory area.

As I have illustrated in details in section 2.4 of my book on abductive cognition (Magnani 2009), first and foremost, anomaly resolution involves the localization of the problem at hand within one or more constituents of the theory, it is then necessary to produce one or more new hypotheses to account for the anomaly, and, finally, these hypotheses need to be evaluated so as to establish which one best satisfies the criteria for theory justification. Hence, anomalies require a change in the theory. We know that empirical anomalies are not alone in generating impasses. The so-called *conceptual problems* represent a particular form of anomaly. Resolving conceptual problems may involve satisfactorily answering questions about the status of theoretical entities: conceptual problems arise from the nature of the claims in the principles or in the hypotheses of the theory. Usually it is necessary to identify the conceptual problem that needs a resolution, for example by delineating how it can concern the adequacy or the ambiguity of a theory, and yet also its incompleteness or (lack of) evidence.

Formal sciences are especially concerned with conceptual problems. The discovery of non-Euclidean geometries presents an interesting case of visual/spatial abductive reasoning, where both *explanatory* and *non-explanatory* aspects are intertwined. First of all it demonstrates a kind of *visual/spatial abduction*, as a strategy for anomaly resolution connected to a form of explanatory and productive visual thinking. Since ancient times the fifth postulate has been held to be not evident. This "conceptual problem" has generated many difficulties about the reliability of the theory of parallels, consisting of the theorems that can be only derived with the help of the fifth postulate. The recognition of this anomaly was crucial to the development of the non-Euclidean revolution. Two thousand years of attempts to resolve the anomaly have produced many fallacious demonstrations of the fifth postulate: a typical attempt was that of trying to prove the fifth postulate from the others. Nevertheless, these attempts have also provided much theoretical speculation about the unicity of Euclidean geometry and about the status of its principles.

Here, I am primarily interested in showing how the anomaly is recognizable. A postulate that is equivalent to the fifth postulate states that for every line l and every point P that does not lie on l, there exists a unique line m through P that is parallel to l. If we consider its model-based (diagrammatic) counterpart (cf. Fig. 2), the postulate may seem "evident" to the reader, but this is because we have been conditioned to think in terms of Euclidean geometry. The definition above represents the most obvious level at which ancient Euclidean geometry was developed as a formal science – a level composed of *symbols* and *propositions*.

Furthermore, when we also consider the other fundamental level, where model-based aspects (diagrammatic) are at play, we can immediately detect

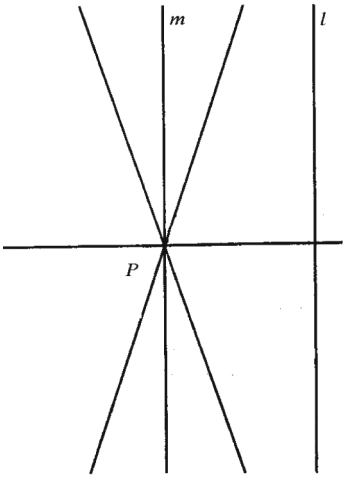


Fig. 2. Diagrammatic counterpart of the Fifth postulate.

a difference between this postulate and the other four if we regard the first principles of geometry as abstractions from experience that we can in turn represent by drawing figures on a blackboard or on a sheet of paper or on our "visual buffer" (Kosslyn and Koenig 1992) in the mind. We have consequently a *double passage* from the sensorial experience to the abstraction (expressed by symbols and propositions) and from this abstraction to the experience (sensorial and/or mental).

We immediately discover that the first two postulates are abstractions from our experiences drawing with a straightedge, the third postulate derives from our experiences drawing with a compass. The fourth postulate is less evident as an abstraction, nevertheless it derives from our measuring angles with a protractor (where the sum of supplementary angles is 180°, so that if supplementary angles are congruent to each other, they must each measure 90°) (Greenberg 1974, p. 17).

In the case of the fifth postulate we are faced with the following serious problems: (1) we cannot verify empirically whether two lines meet, since we can draw only segments, not lines. Extending the segments further and further to find if they meet is not useful, and in fact we cannot continue indefinitely. We are forced to verify parallels indirectly, by using criteria other than the definition; (2) the same holds with regard to the representation in the "limited" visual buffer. The "experience" localizes a problem to solve, an ambiguity, only in the fifth case: in the first four cases our "experience" *verifies* without difficulty the abstraction (propositional and symbolic) itself. In the fifth case the formed images (mental or not) are the images that are able to *explain* the "concept" expressed by the definition of the fifth postulate as problematic (an anomaly): we cannot draw or "imagine" the two lines at infinity, since we can draw and imagine only segments, not the lines themselves.

The chosen visual/spatial image or imagery (in our case the concrete diagram depicted in Fig. 2, derived from the propositional and symbolic level of the definition) plays the role of an explanation of the anomaly previously envisaged in the definition itself. As stated above, the image demonstrates a kind of visual abduction, as a strategy for anomaly localization related to a form of explanatory visual/spatial thinking.

Once the anomaly is detected, the way to anomaly resolution is opened up – in our case, this means that it becomes possible to discover non-Euclidean geometries. That Euclid himself did not fully trust the fifth postulate is revealed by the fact that he postponed using it in a proof for as long as possible – until the twenty-ninth proposition. As is well-known, Proclus tried to solve the anomaly by proving the fifth postulate from the other four. If we were able to prove the postulate in this way, it would become a theorem in a geometry which does not require that postulate (the future "absolute geometry") and which would contain all of Euclid's geometry.

Without showing all the passages of Proclus's argument (Greenberg 1974, p. 119–121) we need only remember that the argument seemed correct because it was proved using a diagram. Yet we now know that we are not allowed to use that diagram to justify a step in a proof. Each step must be proved from stated axioms or previously proven theorems. We may visualize parallel lines as railroad tracks, everywhere equidistant from each other, and the ties of the tracks as being perpendicular to both parallels. Yet this imagery is valid only in Euclidean geometry. In the absence of the parallel postulate we can only consider two lines as "parallel" when, by the definition of "parallel", they do not possess any points in common. It is not possible implicitly to assume that they are equidistant; nor can it be assumed that they have a common perpendicular. This is an example in which a *selected* abduced image is capable of compelling you to make a mistake, and in this way it was used as a means of evaluation in a proof: we have already stated that in this case it is not possible to use that image or imagery to justify a step in a proof because it is not possible to use that image or imagery that attributes to experience more than the experience itself can deliver.

For over two thousand years some of the greatest mathematicians tried to prove Euclid's fifth postulate. For example, Saccheri's strategy for anomaly resolution in the XVIII century was to abduce two opposite hypotheses[6] of the principle, that is, to negate the fifth postulate and derive, using new logical tools coming from non-geometrical sources of knowledge, all theorems from the two alternative hypotheses by trying to detect a contradiction. The aim was indeed that of demonstrating/explaining that the anomaly is simply apparent. We are faced with a kind of non explanatory abduction. New axioms are hypothesized and adopted in looking for outcomes which can possibly help in explaining how the fifth postulate is unique and so not anomalous. At a first sight this case is similar to the case of non-explanatory abduction pointed out in chapter two, section 2.2 of my book on abduction of 2009, speaking of reverse mathematics,

[6] On the strategies adopted in anomaly resolution cf. (Darden 1991, pp. 272–275).

but the similarity is only structural (i.e. guessing "new axioms"): in the case of reverse mathematics axioms are hypothesized to account for already existing mathematical theories and do not aim at explanatory results.[7]

The contradiction in the elliptic case ("hypothesis of obtuse angle", to use the Saccheri's term designing one of the two future elementary non-Euclidean geometries) was found, but the contradiction in the hyperbolic case ("hypothesis of the acute angle") was not so easily discovered: having derived several conclusions that are now well-known propositions of non-Euclidean geometry, Saccheri was forced to resort to a metaphysical strategy for anomaly resolution: "Proposition XXXIII. The 'hypothesis' of acute angle [that is, the hyperbolic case] is absolutely false, because repugnant to the nature of the straight line" (Saccheri (1920)). Saccheri chose to state this result with the help of the somewhat complicated imagery of infinitely distant points: two different straight lines cannot both meet another line perpendicularly at one point, if it is true that all right angles are equal (fourth postulate) and the two different straight lines cannot have a common segment. Saccheri did not ask himself whether everything that is true of ordinary points is necessarily true of an infinitely distant point. In Note II to proposition XXI some "physico-geometrical" experiments to confirm the fifth postulate are also given, invalidated unfortunately by the same incorrect use of imagery that we have observed in Proclus's case. In this way, the anomaly was resolved unsatisfactorily and Euclid was not freed of every fleck: nevertheless, although he did not recognize it, Saccheri had discovered many of the propositions of non-Euclidean geometry (Torretti 1978, p. 48).

In the following sections I will illustrate the example of Lobachevsky's discovery of non-Euclidean geometry where we can see the model-based abductive role played in a discovery process by new considerations concerning visual sense impressions and productive imagery representations.

2.1 Internal and External Representations

Lobachevsky was obliged first of all to rebuild the basic Principles and to this end, it was necessary to consider geometrical principles in a new way, as neither ideal nor *a priori*. New interrelations were created between two areas

[7] Gabbay and Woods (2005) contend – and I agree with them – that abduction *is not intrinsically explanationist*, like for example its description in terms of inference to the best explanation would suggest. Not only, abduction can also be merely *instrumental*. In chapter two of Magnani (2009) I have illustrated various non-explanatory (and instrumental) aspects of abduction and, already in my book on abduction of 2001 (Magnani (2001a)), some examples of abductive reasoning that basically are non-explanatory and/or instrumentalist have been described, without clearly acknowledging it. Gabbay and Woods' distinction between explanatory, non-explanatory and instrumental abduction is orthogonal to mine in terms of the theoretical and manipulative (including the subclasses of sentential and model-based) and further allows us to explore fundamental features of abductive cognition. Hence, if we maintain that E explains E' *only if* the first implies the second, certainly the reverse does not hold. This means that various cases of abduction are consequentialist but not explanationist [other cases are neither consequentialist nor explanationist].

of knowledge: Euclidean geometry and the philosophical tradition of empiricism/sensualism. In the following section I will describe in detail the type of abduction that was at play in this case. Lobachevsky's target is to perform a geometrical abductive process able to create the new and very abstract concept of non-Euclidean parallel lines. The whole epistemic process is mediated by interesting manipulations of external mirror diagrams.

I have already said that for over two thousand years some of the greatest mathematicians tried to prove Euclid's fifth postulate. Geometers were not content to merely construct proofs in order to discover new theorems and thereby to try to resolve the anomaly (represented by its lack of evidence) without trying to reflect upon the status of the symbols of the principles underlying Euclidean geometry. Lobachevsky's strategy for resolving the anomaly of the fifth postulate was first of all to manipulate the symbols, second to rebuild the principles, and then to derive new proofs and provide a new mathematical apparatus; of course his analysis depended on some of the previous mathematical attempts to demonstrate the fifth postulate. The failure of the demonstrations – of the fifth postulate from the other four – that was present to the attention of Lobachevsky, leads him to believe that the difficulties that had to be overcome were due to causes traceable at the level of the first principles of geometry.

We simply can assume that many of the internal visualizations of the working geometers of the past were spatial and imaginary because those mathematicians were precisely operating with diagrams and visualizations. By using internal representations Lobachevsky has to create new external visualizations and to adjust them tweaking and manipulating (Trafton et al. (2005)) the previous ones in some particular ways to generate appropriate spatial transformations (the so-called *geometrical constructions*).[8] In cognitive science many kinds of spatial transformations have been studied, like mental rotation and any other actions to improve and facilitate the understanding and simplification of the problem. It can be said that when a spatial transformation is performed on external visualizations, it is still generating or exploiting an internal representation.

Spatial transformations on external supports can be used to create and transform external diagrams and the resulting internal/mental representations may undergo further mental transformations. Lobachevsky mainly takes advantage of the transformation of external diagrams to create and modify the subsequent internal images. So mentally manipulating both external diagrams and internal representations is extremely important for the geometer that uses both the drawn geometrical figure and her own mental representation. An active role of these external representations, as *epistemic mediators* able to favor scientific discoveries – widespread during the ancient intuitive geometry based on diagrams – can be curiously seen at the beginning of modern mathematics, when new abstract, imaginary, and counterintuitive non-Euclidean entities are discovered and developed.

[8] I maintain that in general spatial transformations are represented by a visual component and a spatial component (Glasgow and Papadias (1992)).

There are *in vivo* cognitive studies performed on human agents (astronomers and physicists) about the interconnection between mental representations and the external scientific visualizations. In these studies "pure" spatial transformations, that is transformations that are performed – and based – on the external visualizations dominate: the perceptual activity seems to be prevalent, and the mental representations are determined by the external ones. The researchers say that there is, in fact, some evidence for this hypothesis: when a scientist mentally manipulates a representation, 71% of the time the source is a visualization, and only 29% of the time it is a "pure" mental representation. Other experimental results show that some of the time scientists seem to create and interpret mental representations that are different from the images in the visual display: in this case it can be hypothesized that scientists use a comparison process to connect their internal representation with the external visualizations (Trafton et al. (2005)).

In general, during the comparison between internal and external representation the scientists are looking for discrepancies and anomalies, but also equivalences and coherent shapes (like in the case of geometers, as we will see below). The comparison between the transformations acted on external representations and their previously represented "internal" counterpart forces the geometer to merge or to compare the two sides (some aspects of the diagrams correspond to information already represented internally as symbolic-propositional).[9]

External geometrical diagrams activate perceptual operations, such as searching for objects that have a common shape and inspecting whether three objects lie on a straight line. They contain permanent and invariant geometrical information that can be immediately perceived and kept in memory without the mediation of deliberate inferences or computations, such as whether some configurations are spatially symmetrical to each other and whether one group of entities has the same number of entities as another one. Internal operations prompt other cognitive operations, such as making calculations to get or to envision a result. In turn, internal representations may have information that can be directly retrieved, such as the relative magnitude of angles or areas.

3 Mirror Diagrams and the Infinite

As previously illustrated, the failure of the demonstrations (of the fifth postulate from the other four) of his predecessors induced Lobachevsky to believe that the difficulties that had to be overcome were due to causes other than those which had until then been focused on.

[9] Usually scientists try to determine *identity*, when they make a comparison to determine the individuality of one of the objects; *alignment*, when they are trying to determine an estimation of fit of one representation to another (e.g. visually inspecting the fit of a rough mental triangular shape to an external constructed triangle); and *feature comparison*, when they compare two things in terms of their relative features and measures (size, shape, color, etc.) (Trafton et al. (2005)).

Lobachevsky was obliged first of all to rebuild the basic principles: to this end, it was necessary to consider geometrical principles in a new way, as neither ideal nor *a priori*. New interrelations were created between Euclidean geometry and some claims deriving from the philosophical tradition of empiricism/sensualism.

3.1 Abducing First Principles Through Bodily Contact

From this Lobachevskyan perspective the abductive attainment of the basic concepts of any science is in terms of senses: the basic concepts are always acquired through our *sense impressions*. Lobachevsky builds geometry upon the concepts of body and bodily contact, the latter being the only "property" common to all bodies that we ought to call geometrical. The well-known concepts of depthless surface, widthless line and dimensionless point were constructed considering different possible kinds of bodily contact and dispensing with, *per abstractionem*, everything but preserving the contact itself: these concepts "[...] exist only in our representation; whereas we actually measure surfaces and lines by means of bodies" for "[...] in nature there are neither straight lines nor curved lines, neither plane nor curved surfaces; we find in it only bodies, so that all the rest is created by our imagination and exists just in the realm of theory" (Lobachevsky 1897, Introduction). The only thing that we can know in nature is movement "[...] without which sense impressions are impossible. Consequently all other concepts, e.g. geometrical concepts, are generated artificially by our understanding, which derives them from the properties of movement; this is why space in itself and by itself does not exist for us" (Lobachevsky 1897).

It is clear that in this inferential process Lobachevsky performs a kind of *model-based* abduction, where the perceptual role of sense impressions and their experience with bodies and bodily contact is cardinal in the generation of new concepts. The geometrical concepts are "[...] generated artificially by our understanding, which derives them from the properties of movement". Are these abductive hypotheses explanatory or not? I am inclined to support their basic "explanatory" character: they furnish an explanation of our sensorial experience with bodies and bodily contact in ideal and abstract terms.

On the basis of these foundations Lobachevsky develops the so-called *absolute geometry*, which is independent of the fifth postulate: "Instead of commencing geometry with the plane and the straight line as we do ordinarily, I have preferred to commence it with the sphere and the circle, whose definitions are not subject to the reproach of being incomplete, since they contain the generation of the magnitudes which they define" (Lobachevsky 1929, p. 361).

This leads Lobachevsky to abduce a very remarkable and modern hypothesis – anticipatory of the future Einstein's theoretical atmosphere of general relativity – which I consider to be largely *image-based:* since geometry is not based on a perception of space, but constructs a concept of space from an experience of bodily movement produced by physical forces, there could be place in science for two or more geometries, governing different kinds of natural forces:

> To explain this idea, we assume that [...] attractive forces decrease because their effect is diffused upon a spherical surface. In ordinary Geometry the area of a spherical surface of radius r is equal to $4r^2$, so that the force must be inversely proportional to the square of the distance. In Imaginary Geometry I found that the surface of the sphere is
>
> $$(e^r - e^{-r})^2,$$
>
> and it could be that molecular forces have to follow that geometry [...]. After all, given this example, merely hypothetical, we will have to confirm it, finding other more convincing proofs. Nevertheless we cannot have any doubts about this: forces by themselves generate everything: movement, velocity, time, mass, matter, even distances and angles (Lobachevsky 1897, p. 9).

Lobachevsky did not doubt that something, not yet observable with a microscope or analyzable with astronomical techniques, accounted for the reliability of the new non-Euclidean imaginary geometry. Moreover, the principles of geometry are held to be testable and it is possible to prepare an experiment to test the validity of the fifth postulate or of the new non-Euclidean geometry, the so-called *imaginary geometry*. He found that the defect of the triangle formed by Sirius, Rigel and Star No. 29 of Eridanus was equal to $3.727 + 10^{-6}$ seconds of arcs, a magnitude too small to be significant as a confirmation of imaginary geometry, given the range of observational error. Gauss too had claimed that the new geometry might be true on an astronomical scale. Lobachevsky says:

> Until now, it is well-known that, in Geometry, the theory of parallels had been incomplete. The fruitlessness of the attempts made, since Euclid's time, for the space of two thousand years, aroused in me the suspicion that the truth, which it was desired to prove, was not contained in the data themselves; that to establish it the aid of experiment would be needed, for example, of astronomical observations, as in the case of other laws of nature. When I had finally convinced myself of the justice of my conjecture and believed that I had completely solved this difficult question I wrote, in 1826, a memoir on this subject *Exposition succincte des principes de la Géométrie* (Lobachevsky 1897, p. 5).

With the help of the explanatory abductive role played by the new sensualist considerations of the basic principles, by the empiricist view and by a very remarkable productive visual hypothesis, Lobachevsky had the possibility to proceed in discovering the new theorems. Following Lobachevsky's discovery the fifth postulate will no longer be considered in any way anomalous – we do not possess any proofs of the postulate, because this proof is *impossible*. Moreover, the new non-Euclidean hypothesis is reliable: indeed, to understand visual thinking we have also to capture its status of guaranteeing the reliability of a hypothesis. In order to prove the relative consistency of the new non-Euclidean geometries we should consider some very interesting visual and mathematical

"models" proposed in the second half of XIX century (i.e. the Beltrami-Klein and Poincaré models), which involve new uses of visual images in theory assessment.

In summary, the abductive process of Lobachevsky's discovery can be characterized in the following way, taking advantage of the nomenclature I have introduced in chapter two of my book on abduction of 2009 (Magnani (2009)):

1. the inferential process Lobachevsky performs to rebuild the first principles of geometry is prevalently a kind of *manipulative* and *model-based* abduction, endowed with an *explanatory* character: the new abduced principles furnish an explanation of our sensorial experience with bodies and bodily contact in ideal and abstract terms;
2. at the same time the new principles found offer the chance of further *multimodal*[10] and *distributed* abductive steps (that is based on both visual and sentential aspects, and on both internal and external representations) which are mainly *non-explanatory* and provide unexpected mathematical results. These further abductive processes:
   a. first of all have to provide a different multimodal way of describing parallelism (both from a diagrammatical and propositional perspective, cf. Subsect. 3.4 and Fig. 5);
   b. second, on the basis of the new concept of parallelism it will be possible to derive new theorems of a new non-Euclidean geometrical system exempt from inconsistencies just like the Euclidean system. Of course this process shows a moderately instrumental character, more or less present in all abductions (cf. below Sect. 4).

Let us illustrate how Lobachevsky continues to develop the absolute geometry. The immediate further step is to define the concept of plane, which is defined as the geometrical locus of the intersections of equal spheres described around

[10] Multimodality of abduction depicts hybrid aspects of abductive reasoning. Thagard (2005, 2007) observes that abductive inference can be visual as well as verbal, and consequently acknowledges the sentential, model-based, and manipulative nature of abduction. Moreover, both data and hypotheses can be visually represented and it is an interesting question whether hypotheses can be represented using all sensory modalities (cf. also (Magnani 2009, chapter four)). For vision the answer is obvious, as images and diagrams can clearly be used to represent events and structures that have causal effects, but hypotheses can be also represented using other sensor modalities:

> I may recoil because something I touch feels slimy, or jump because of a loud noise, or frown because of a rotten smell, or gag because something tastes too salty. Hence in explaining my own behavior my mental image of the full range of examples of sensory experiences may have causal significance. Applying such explanations of the behavior of others requires projecting onto them the possession of sensory experiences that I think are like the ones that I have in similar situations. [. . .] Empathy works the same way, when I explain people's behavior in a particular situation by inferring that they are having the same kind of emotional experience that I have in similar situations (Thagard (2007)).

.

two fixed points as centers, and, immediately after, the concept of straight line (for example BB' in the mirror diagram of the Fig. 3) as the geometrical locus of the intersections of equal circles, all situated in a single plane and described around two fixed points of this plane as centers. The straight line is so defined by means of "finite" parts (segments) of it: we can prolong it by imaging a repeatable movement of rotation around the fixed points (cf. Fig. 3) (Lobachevsky 1838, §25).

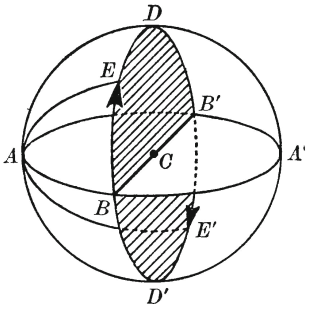


Fig. 3. The concept of straight line defined as the geometrical locus of the intersections of equal spheres described around two fixed points as centers (example of the use of a mirror diagram).

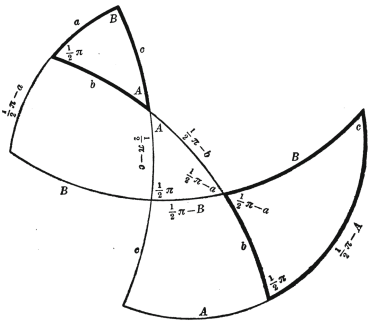


Fig. 4. Example of the exploitation of the analogy plane/spherical surface by means of a diagram that exploits the perspective of the two-dimensional flat plane.

Rectilinear angles (which express arcs of circle) and dihedral angles (which express spherical lunes) are then considered; and the solid angles too, as generic parts of spherical surfaces – and in particular the interesting spherical triangles. π means for Lobachevsky the length of a semicircumference, but also the solid angle that corresponds to a semisphere (straight angle). The surface of the

spherical triangles is always less than π, and, if π, coincides with the semisphere. The theorems about the perpendicular straight lines and planes also belong to absolute geometry.

3.2 Expansion of Scope Strategy

We have to note some general cognitive and epistemological aspects which characterize the development of this Lobachevskyan absolute geometry.

Spherical geometry is always treated together with the plane geometry: the definitions about the sphere are derived from the ones concerning the plane when we substitute the straight lines (geodetics in the plane) with the maximal circles (geodetics in the spherical surface). Lobachevsky says that the maximal circle on the sphere with respect to the other circles presents "properties" that are "similar" to the ones belonging to the straight line with respect to all the segments in the plane (Lobachevsky 1838, §66). This is an enhancement, by means of a kind of *analogical* reasoning, reinforced by the external mirror diagrams, of the internal representation of the concept of straight line. The straight line can be in some sense thought (because it is "seen" and "imagined" in the various configurations provided by the external diagrams) as "belonging" to various types of surfaces, and not only to the plane. Consequently, mirror diagrams not only manage consistency requirements, they can also act in an additive way, providing new "perspectives" and information on old entities and structures. The directly perceivable information strongly guides the discoverer's selections of moves by servicing the discovery strategy *expansion-of-the-scope* (of the concept of straight line). This possibility was not indeed available at the simple level of the internal representation. The Fig. 4 (Lobachevsky 1838, §79) is another example of the exploitation of the analogy plane/spherical surface by means of a diagram that exploits the perspective of the two-dimensional flat plane.

3.3 Infinite/Finite Interplay

In all the previous cases the external representations are constructions that have to respect the empirical attitude described above: because of the fact that the geometrical bodies are characterized by their "finiteness" the external representation is just a coherent mirror of finite internal images. The "infinite" can be perceived in the "finite" constructions because the infinite is considered only as something potential that can be just mentally and artificially thought: "defined artificially by our understanding". As the modern axiomatic method is absent, the geometer has to conceptualize infinite situations exploiting the finite resources offered by diagrams. In front of the question: "How is it that the finite human resources of internal representations of human mind can conceptualize and formalize abstract notions of infinity?" – notions such as the specific ones embedded in the non-Euclidean assumptions – the geometer is aware we perceive a finite world, act upon it, and think about it. Moreover, the geometer operates in "[...] a combination of perceptual *input*, physical *output*, and internal mental processes. All three are finite. But by thinking about the possibility

of performing a process again and again, we can easily reach out towards the potential infinite" (Tall (2001)). Lobachevsky states: "Which part of the lines we would have to disregard is arbitrary", and adds, "our senses are deficient" and it is only by means of the "artifice" consisting of the continuum "enhancement of the instruments" that we can overcome these limitations (Lobachevsky 1838, §38). Given this epistemological situation, it is easy to conclude saying that *instruments* are not just and only telescopes and laboratory tools, but also diagrams.

Let us continue to illustrate the geometer's inventions. In the Proposition 27 (a theorem already proved by Euler and Legendre) of the *Geometrical Researches of the Theory of Parallels*, published in 1840 (Lobachevsky 1891), Lobachevsky states that if A, B, and C are the angles of a spherical triangle, the ratio of the area of the triangle to the area of the sphere to which it belongs will be equal to the ratio of

$$\frac{1}{2}(A + B + C - \pi)$$

to four right angles; that the sum of the three right angles of a rectilinear triangle can never surpass two right angles (Prop. 19), and that, if the sum is equal to two right angles in any triangle, it will be so in all (Prop. 20).

3.4 Non-Euclidean Parallelism: Coordination and Inconsistency Detection

The basic unit is the manipulation of diagrams. Before the birth of the modern axiomatic method the geometers still and strongly have to exploit external diagrams, to enhance their thoughts. It is impossible to mental imaging and evaluating the alternative sequences of symbolic calculations being only helped by the analytic tools, such as various written equations and symbols and marks: it is impossible to do a complete anticipation of the possible outcomes, due to the limited power of working memory and attention. Hence, because of the complexity of the geometrical problem space and the limited power of working memory, complete mental search is impossible or difficult. Geometers may use perceptual external biases to make decisions. Moreover, in those cognitive settings, lacking in modern axiomatic theoretical awareness, certainly perceptual operations were epistemic mediators which need less attentional and working memory resources than internal operations. "The directly perceived information from external representations and the directly retrieved information from internal representation may elicit perceptual and cognitive biases, respectively, on the selections of actions. If the biases are inconsistent with the task, however, they can also misguide actions away from the goal. Learning effect can occur if a task is performed more than once. Thus, the decision on actions can also be affected by learned knowledge" (Zhang 1997, p. 186).

The new external diagram proposed by Lobachevsky (the diagram of the drawn parallel lines of Fig. 5) (Lobachevsky 1891) is a kind of *analogous* both

of the mental image we depict in the mental visual buffer and of the symbolic-propositional level of the postulate definition. It no longer plays the *explanatory* role of showing an anomaly, like it was in the case of the diagram of Fig. 2 (and of other similar diagrams) during the previous centuries. I have already said I call this kind of external tool in the geometrical reasoning *mirror diagram*. In general this diagram mirrors the internal imagery and provides the possibility of detecting anomalies, like it was in the case of the similar diagram of Fig. 2. The external representation of geometrical structures often activates direct perceptual operations (for example, identify the parallels and search for the limits) to elicit consistency or inconsistency routines. Sometimes the mirror diagram biases are inconsistent with the task and so they can make the task more difficult by misguiding actions away from the goal. If consistent, we have already said that they can make the task easier by instrumentally and non-explanatorily guiding actions toward the goal. In certain cases the mirror diagrams biases are irrelevant, they should have no effects on the decision of abductive actions, and play lower cognitive roles.

In the case of the diagram of the parallel lines of the similar Fig. 2, it was used in the history of geometry to make both consistent and in-consistent the fifth Euclidean postulate and, consequently, the new non-Euclidean perspective (more details on this epistemological situation are given in (Magnani 2001b)).

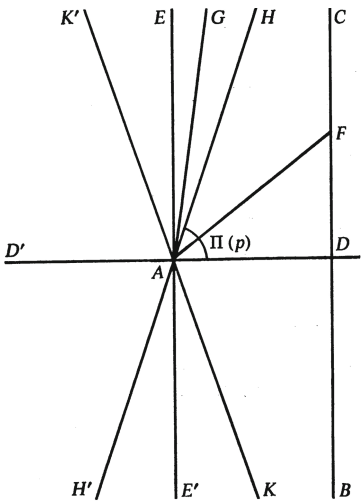


Fig. 5. Non-Euclidean parallel lines.

I said that in some cases the mirror diagram plays a negative role and inhibits further creative abductive theoretical developments. As I have already indicated (p. 7), Proclus tried to solve the anomaly by proving the fifth postulate from the other four. If we were able to prove the postulate in this way, it would become a

theorem in a geometry which does not require that postulate (the future "absolute geometry") and which would contain all of Euclid's geometry. We need only remember that the argument seemed correct because it was proved using a diagram. In this case the mirror diagram biases were consistent with the task of *justifying* Euclidean geometry and they made this task easier by guiding actions toward the goal, but they inhibited the discovery of non-Euclidean geometries (Greenberg 1974, pp. 119–121); cf. also (Magnani 2001b, pp. 166–167).

In sum, contrarily to the diagram of Fig. 2, the diagram of Fig. 5 does not aim at explaining anything given, it is fruit of a non-explanatory and instrumental abduction, as I have anticipated at p. 13: the new related principle/concept of parallelism offers the chance of further *multimodal* and *distributed* abductive steps (based on both visual and sentential aspects, and on both internal and external representations) which are mainly *non-explanatory*. On the basis of the new concept of parallelism it will be possible to derive new theorems of a new non-Euclidean geometrical system exempt from inconsistencies just like the Euclidean system (cf. below Sect. 4).

The diagram now favors the new definition of parallelism (Lobachevsky 1891, Prop. 16), which introduces the non-Euclidean atmosphere: "All straight lines which in a plane go out from a point can, with reference to a given straight line in the same plane, be divided in two classes – into *cutting* and *not-cutting*. The boundary lines of the one and the other class of those lines will be called *parallel to the given lines*" (p. 13).

The external representation is easily constructed like in Fig. 5 of (Lobachevsky 1891, p. 13), where the angle HAD between the parallel HA and the perpendicular AD is called the angle of parallelism, designated by $\Pi(p)$ for $AD = p$. If $\Pi(p)$ is $< \frac{1}{2}\pi$, then upon the other side of AD, making the same angle $DAK = \Pi(p)$ will lie also a line AK, parallel to the prolongation DB of the line DC, so that under this assumption we must also make a distinction of sides in parallelisms. Because of the fact that the diagrams can contemplate only finite parts of straight lines it is easy to represent this new postulate in this mirror image: we cannot know what happens at the infinite neither in the internal representation (because of the limitations of visual buffer), nor in the external representation: "[. . .] in the uncertainty whether the perpendicular AE is the only line which does not meet DC, we will assume it may be possible that there are still other lines, for example AG, which do not cut DC, how far so ever they may be prolonged" (Lobachevsky 1891). So the mirror image in this case is seen as consistently supporting the new non-Euclidean perspective. The idea of constructing an external diagram of a non-Euclidean situation is considered normal and reasonable. The diagram of Fig. 5 is now exploited to "unveil" new fruitful consequences.

A first analysis of the exploitation of what I call *unveiling diagrams* in the discovery of the notion of non-Euclidean parallelism is presented in the following section related to the exploitation of diagrams at the stereometric level.[11]

[11] Magnani and Dossena (2005); Dossena and Magnani (2007) illustrate that external representations like the ones I call *unveiling diagrams* can enhance the consistency of a cognitive process but also provide more radically creative suggestions for new useful information and discoveries.

4 Unveiling Diagrams in Lobachevsky's Discovery as Gateways to Imaginary Entities

4.1 Euclidean/Non-Euclidean Model Matching Strategy

Lobachevsky's target is to perform a geometrical abductive process able to create new and very abstract entities: the whole epistemic process is mediated by interesting manipulations of external unveiling diagrams. The first step toward the exploitation of what I call *unveiling diagrams* is the use of the notion of non-Euclidean parallelism at the stereometric level, by establishing relationships between straight lines and planes and between planes: Proposition 27 (already proved by Lexell and Euler): "*A three-sided solid angle equals the half sum of surface angles less a right-angle*" (p. 24, Fig. 6). Proposition 28 (directly derived from Prop. 27): "*If three planes cut each other in parallel lines, then the sum of the three surface angles equals two rights*" (p. 28), Fig. 7. These achievements are absolutely important: it is established that for a certain geometrical configuration of the new geometry (the three planes cut each other in parallel lines that are parallel in Lobachevskyan sense) some properties of the ordinary geometry hold.

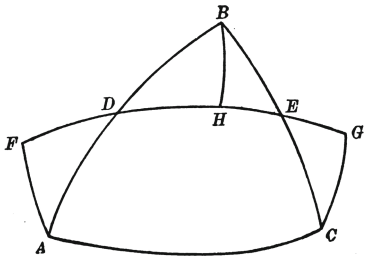


Fig. 6. A three-sided solid angle equals the half sum of surface angles less a right-angle.

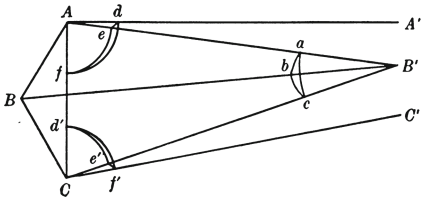


Fig. 7. If three planes cut each other in parallel lines, then the sum of the three surface angles equals two rights.

The important notions of *oricycle* and *orisphere* are now defined to search for a possible symbolic counterpart able to express a foreseen consistency (as a

justification) of the non-Euclidean theory. This consistency is looked at from the point a view of a possible "analytic" solution, that is in terms of verbal-symbolic (not diagrammatic) results (equations).

Such is the case of the Proposition 31. "*We call boundary line (oricycle) that curve lying in a plane for which all perpendiculars erected at the mid-points of chords are parallel to each other.* [...] The perpendicular DE erected upon the chord AC at its mid-point D will be parallel to the line AB, which we call the *Axis of the boundary line*" (pp. 30–31), cf. Fig. 8. Proposition 34. "*Boundary surface*[12] we call that surface which arises from the revolution of the boundary line about one of its axes, which, together with all other axes of the boundary-line, will be also an axis of the boundary surface" (p. 33). Moreover, the intersections of the orisphere by its diametral planes are limit circles. The limit circle arcs are called the sides, and the dihedral angles between the planes of these arcs the angles of the "orisphere triangle".

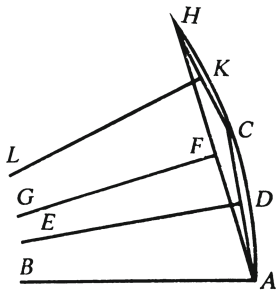


Fig. 8. Oricycle: the curve lying in a plane for which all perpendiculars erected at the mid-points of chords are parallel to each other. The perpendicular DE erected upon the chord AC at its mid-point D will be parallel to the line AB, which is called the axis of the boundary line.

A part of the surface of the orisphere bounded by three limit circle arcs will be called an orisphere triangle. From Prop. 28 follows that the sum of the angles of an orisphere triangle is always equal to two right angles: "everything that is demonstrated in the ordinary geometry of the proportionality of the sides or rectilinear triangles can therefore be demonstrated in the same manner in the pangeometry"[13] (Lobachevsky 1929, p. 364) of the orisphere triangles if only we will replace the lines parallel to the sides of the rectilinear triangle by orisphere arcs drawn through the points of one of the sides of the orisphere triangle and all making the same angle with this side. To conclude, the orisphere is a "partial" *model* of the Euclidean plane geometry.

The last constructions of the Lobachevskyan abductive process give rise to two fundamental *unveiling diagrams* (cf. Figs. 9 and 11) that accompany the

[12] Also called limit sphere or orisphere.

[13] Lobachevsky called the new theory "imaginary geometry" but also "pangeometry".

remaining proofs. They are more abstract and exploit "audacious" representations in the perspective of three dimensional geometrical shapes.

The construction given in Fig. 9 aims at diagrammatically "representing" a stereometric non-Euclidean form built on a *rectilinear right angled triangle* ABC to which the Theorem 28 above can be applied (indeed the parallels AA', BB', CC', which lie on the three planes are parallels in non-Euclidean sense), so that Lobachevsky is able to further apply symbolic identifications; the planes make with each other the angles $\Pi(a)$ at AA', a right angle at CC', and, consequently $\Pi(a')$ at BB'.[14] The diagram is enhanced by constructing a spherical triangle mnk, in which the sides are $mn = \Pi(c)$, $kn = \Pi(\beta)$, $mk = \Pi(a)$ and the opposite angles are $\Pi(a)$, $\Pi(\alpha')$, $\frac{1}{2}\pi$ realizing that with the "existence" of a rectilinear triangle with the sides a, b, c (like in the case of the previous one) "we must admit" the existence of a related spherical triangle (cf. Fig. 10), etc. Not only, a *boundary surface* (orisphere) can be constructed, that passes through the point A with AA' as axis, and those intersections with the planes the parallels form a boundary triangle (that is a triangle situated upon the given orisphere), whose sides are $B''C'' = p$, $C''A = q$, $B''A = r$, and the angles opposite to them $\Pi(\alpha)$, $\Pi(\alpha')$, $\frac{1}{2}\pi$ and where consequently (this follows from the Theorem 34):

$$p = r \sin \Pi(a), q = r \cos \Pi(a).$$

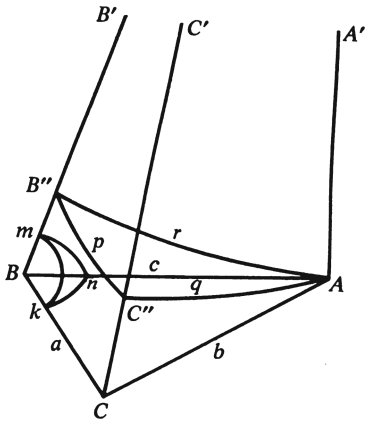


Fig. 9. Unveiling diagram. Diagram that represents a stereometric non-Euclidean form built on a *rectilinear right angled triangle* ABC to which the Theorem 28 can be applied (indeed the parallels AA', BB', CC', which lie on the three planes are parallels in non-Euclidean sense).

[14] Given that Lobachevsky designates the size of a line by a letter with an accent added, e.g. x', in order to indicate this has a relation to that of another line, which is represented by the same letter without the accent x, "which relation is given by the equation $\Pi(x) + \Pi(x') = \pi$" (Prop. 35).

As I will illustrate in the following subsections, in this way Lobachevsky is able to further apply symbolic identifications and to arrive to new equations which consistently (and at the same time) connect Euclidean and non-Euclidean perspectives. This kind of diagram strongly guides the geometer's selections of moves by eliciting what I call the *Euclidean-inside non-Euclidean* "model matching strategy".

Inside the perspective representations (given by the fundamental unveiling diagram of a *non-Euclidean structure*, cf. Fig. 9), a *Euclidean spherical triangle* and the *orisphere* (and its boundary triangle where the Euclidean properties hold) are constructed. The directly perceivable information strongly guides the geometer's selections of moves by eliciting the Euclidean-inside non-Euclidean "model matching strategy" I have quoted above. This maneuver also constitutes an important step in the affirmation of the modern "scientific" concept of *model*. We have to note that other perceptions activated by the diagram are of course disregarded as irrelevant to the task, as it usually happens when exploiting external diagrammatic representations in reasoning processes. Because not everything in external representations is always relevant to a task, high level cognitive mechanisms need to use task knowledge (usually supplied by task instructions) to direct attention and perceptual processes to the relevant features of external representations.

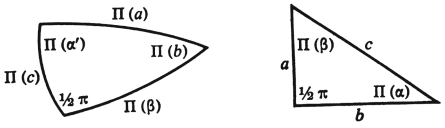


Fig. 10. Spherical triangle and rectilinear triangle.

The different selected representational system – that still uses Euclidean icons – determines in this case quite different possibilities of constructions, and thus different results from iconic experimenting. New results are derived in diagrammatic reasoning through modifying the representational systems, adding new meaning to them, or in reconstructing their systematic order.

4.2 Consistency-Searching Strategy

This external representation in terms of the unveiling diagram illustrated in Fig. 9 activates a perceptual reorientation in the construction (that is identifies possible further constructions); in the meantime the consequent new generated internal representation of the external elements activates directly retrievable information (numerical values) that elicits the strategy of building further non-Euclidean structures together with their *analytic counterpart* (cf. below the non-Euclidean trigonometry equations). Moreover, the internal representation of the stereometric figures activates cognitive operations related to the *consistency-searching strategy*. In this process, new "imaginary" and strange mathematical

entities, like the oricycle and the orisphere, are – non-explanatorily – abduced and *unveiled*, and related to ordinary and unsuspected perceptive entities.

Finally, it is easy to identify in the proof the differences between perceptual and other cognitive operations and the differences between *sequential* – the various steps of the constructed unveiling diagram – and *parallel* perceptual operations. Similarly, it is easy to distinguish between the forms that are directly perceptually inspected and the elements that are mentally computed or computed in external symbolic configurations.

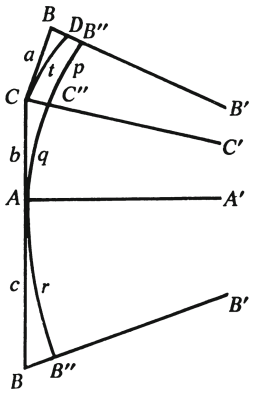


Fig. 11. A final productive unveiling diagram.

To arrive to the second unveiling diagram the old diagram (cf. Fig. 9) is further enhanced by a new construction, by breaking the connection of the three principal planes along the line BB', and by turning them out from each other so that they, together with all the lines lying in them, come to lie in one plane, where consequently the arcs p, q, r will unite to a single arc of a boundary-line (oricycle). This goes through the point A and has AA' as its axis, in such a manner that (Fig. 11) on the one side will lie, the arcs q and p, the side b of the triangle, which is perpendicular to AA' at A, the axis CC' going from the end of b parallel to AA' and through C'' the union point of p and q, the side a perpendicular to CC' at the point C, and from the end-point of a the axis BB' parallel to AA' which goes through the end-point B'' of the arc p, etc. Finally, taking CC' as axis, a new boundary line (an arc of oricycle) from the point C to its intersection with the axis BB' is constructed. What happens?

4.3 Loosing Intuition

In this case we see that the external representation completely looses its spatial intuitive interest and/or its capacity to simulate internal spatial representations: it is not useful to represent it as an internal spatial model in order to enhance the

problem solving activity. The diagram of Fig. 11 does not have to depict internal forms coherent from the intuitive spatial point of view, it is just devoted to suitably "unveil" the possibility of further calculations by directly activating perceptual information that, in conjunction with the non-spatial information and cognitive operations provided by internal representations in memory, determine the subsequent problem solving behavior. This diagram does not have to prompt an internal "spatially" intuitively coherent model. Indeed *perception* often plays an autonomous and central role, it is not a peripheral device. In this case the end product of perception and motor operations coincides with the intermediate data highly analyzed, processed, and transformed, that is prepared for high-level cognitive mechanisms in terms of further *analytic* achievements (the equations).[15]

We have to note that of course it cannot be said that the external representation would work independently without the support of anything internal or mental. The mirror and unveiling diagrams have to be processed by perceptual mechanisms that are of course internal. And in this sense the end product of the perceptual mechanisms is also internal. But it is not an internal model of the external representation of the task: the internal representation is the knowledge and structure of the task in memory; and the external representation is the knowledge and structure of the task in the environment. The end product of perception is merely the situational information in working memory that usually only reflects a fraction (crucial) of the external representation (Zhang 1997). At this point it is clear that the perceptual operations generated by the external representations "mediated" by the unveiling diagrams are central as mechanisms of the whole geometrical abductive and manipulative process; they are not less fundamental than the cognitive operations activated by internal representations, in terms of images and/or symbolic-propositional. They constitute an extraordinary example of complex and perfect coordination between perceptual, motor, and other inner cognitive operations.

Let us conclude the survey on Lobachevsky's route to an acceptable assessment of its non-Euclidean theory. By means of further *symbolic/propositional* designations taken from both internal representations followed from previous results and "externalized" calculations, the reasoning path is constrained to find a general "analytic" counterpart for (some aspects of) the non-Euclidean geometry (we skip the exposition of this complicated passage – cf. (Lobachevsky 1891)). Therefore we obtain the equations

$$\sin \Pi(c) = \sin \Pi(a) \sin \Pi(b)$$
$$\sin \Pi(\beta) = \cos \Pi(\alpha) \sin \Pi(a)$$

Hence we obtain, by mutation of the letters,

$$\sin \Pi(\alpha) = \cos \Pi(\beta) \sin \Pi(b)$$

[15] In other problem solving cases, the end product of perception – directly picked-up – is the end product of the whole problem solving process.

$$\cos \Pi(b) = \cos \Pi(c) \cos \Pi(\alpha)$$
$$\cos \Pi(a) = \cos \Pi(c) \cos \Pi(\beta)$$

that express the mutual dependence of the sides and the angles of a non-Euclidean triangle. In these equations of plane non-Euclidean geometry we can pass over the equations for spherical triangles. If we designate in the right-angled spherical triangle (Fig. 10) the sides $\Pi(c)$, $\Pi(\beta)$, $\Pi(a)$, with the opposite angles $\Pi(b)$, $\Pi(\alpha')$, by the letters a, b, c, A, B, then the obtained equations take of the form of those which we know as the equations of spherical trigonometry for the right-angled triangle

$$\sin(a) = \sin(c) \sin(A)$$
$$\sin(b) = \sin(c) \sin(B)$$
$$\cos(A) = \cos(A) \sin(B)$$
$$\cos(B) = \cos(B) \sin(A)$$
$$\cos(c) = \cos(a) \cos(b).$$

The equations are considered to "[...] attain for themselves a sufficient foundation for considering the assumption of imaginary geometry as possible" (p. 44). The new geometry is considered exempt from possible inconsistencies together with the acknowledgment of the reassuring fact that it presents a very complex system full of surprisingly harmonious conclusions. A new contradiction which could have emerged and which would have forced to reject the principles of the new geometry would have been already contained in the equations above. Of course this is not true from the point of view of modern deductive axiomatic systems and a satisfactory model of non-Euclidean geometry has not yet been built (as Beltrami and Klein will do with the so-called "Euclidean models of non-Euclidean geometry").[16] As for now the argument rests on a formal agreement between two sets of equations, one of which is derived from the new non-Euclidean geometry. Moreover, the other set of equations does not pertain to Euclidean geometry; rather they are the equations of spherical trigonometry that does not depend on the fifth postulate (as maintained by Lobachevsky himself). Nevertheless, we can conclude that Lobachevsky is not far from the modern idea of scientific *model*.

We can say that geometrical diagrammatic thinking represented the capacity to extend *finite* perceptual experiences to give known (Euclidean) and infinite unknown (non-Euclidean) mathematical structures that appear consistent in themselves and that have quite different properties each other.

Many commentators (and myself in Magnani (2001b)) contend that Kant did not imagine that non-Euclidean concepts could in some way be constructed in

[16] On the limitations of the Lobachevskyan perspective cf. Torretti (1978) and Rosenfeld (1988).

intuition[17] (a Kantian expression which indicated our iconic external representation), through the mediation of a model, that is preparing and constructing a Euclidean model of a specific non-Euclidean concept (or group of concepts). Yet Kant also wrote that "[...] the use of geometry in natural philosophy would be insecure, unless the notion of space is originally given by the nature of the mind (so that if anyone tries to frame in his mind any relations different from those prescribed by space, he will labor in vain, for he will be compelled to use that very notion in support of his figment)" (Kant 1968, section 15E).

(Torretti 2001, p. 160) observes:

> I find it impossible to make sense of the passage in parentheses unless it refers precisely to the activity of constructing Euclidean models of non-Euclidean geometries (in a broad sense). We now know that one such model (which we ought rather to call quasi-Euclidean, for it would represent plane Lobachevskian geometry on a sphere with radius $\sqrt{-1}$) is mentioned in the *Theorie der Parallellinien* that Kant's fellow Königsbergian Johann Heinrich Lambert (1786) wrote about 1766. There is no evidence that Kant ever saw this tract and the few extant pieces of his correspondence with Lambert do not contain any reference to the subject, but, in the light of the passage I have quoted, it is not unlikely that Kant did hear about it, either from Lambert himself, or from a shared acquaintance, and raised the said objection.

I agree with Torretti, Kant had a very wide perspective about the resources of "intuition", anticipating that a geometer would have been "compelled" to use the notion of space "given by nature", that is the one that is at the origins of our external representation, "in support of his figment", for instance the non-Euclidean Lobachevskyan abstract structures we have treated above in Fig. 9, which exhibits the non-Euclidean through the Euclidean.

5 Conclusion

The analysis of mirror and unveiling diagrams described in this article, taking advantage of the cognitive-epistemological reconstruction of the discovery of non-Euclidean geometry, entails some general consequences concerning the epistemology of mathematics and formal sciences. In our case the concept of mirror diagram plays a fundamental explanatory role in the epistemological removal of obstacles and obscurities related to the ambiguities of the problem of parallel lines and, in general, in enhancing mathematical knowledge regarding critical situations. In the case of the more instrumental unveiling diagrams, the allocating and switching of attention between internal and external representations better reveals how to govern the reasoning strategy at hand by integrating them in a more dynamic and complicated way. I have also illustrated how important

[17] We have seen how Lobachevsky did this by using the Fig. 9.

abductive heuristics play a fundamental role in these creative epistemic endeavors: expansion of scope strategy, Euclidean/non-Euclidean model matching strategy, consistency-searching strategy. My account in terms of mirror and unveiling diagrams seems empirically adequate to integrate findings from research on cognition and findings from historical-epistemological research into models of actual mathematical practices. I contend that the assessment of the fit between cognitive findings and historical-epistemological practices helps to elaborate richer and more realistic models of cognition and presents a significant advance over previous epistemological work on actual mathematical reasoning and practice.

Acknowledgements. Parts of this article were originally published in chapter two of L. Magnani *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, Springer, Heidelberg, 2009.

References

Darden L (1991) Theory change in science: strategies from mendelian genetics. Oxford University Press, Oxford

Dossena R, Magnani L (2007) Mathematics through diagrams: microscopes in non-standard and smooth analysis. In: Magnani L, Li P (eds) Model-based reasoning in science, technology, and medicine. Springer, Heidelberg, pp 193–213

Gabbay DM, Woods J (2005) The reach of abduction. North-Holland, Amsterdam

Glasgow JI, Papadias D (1992) Computational imagery. Cogn Sci 16:255–394

Greenberg MJ (1974) Euclidean and non-euclidean geometries. Freeman and Company, New York

Hutchins E (1995) Cognition in the wild. The MIT Press, Cambridge

Kant I (1929) Critique of pure reason. MacMillan, London. (Trans: Kemp Smith N) originally published 1787, reprint 1998

Kant I (1968) Inaugural dissertation on the forms and principles of the sensible and intelligible world (1770). In: Kerferd G, Walford D (eds) Kant. Selected Pre-Critical Writings, pp 45–92. Manchester University Press, Manchester. (Trans: Kerferd GB, Walford DE, Handyside J, Kant I) Kant's Inaugural Dissertation and Early Writings on Space, pp 35-85. Open Court, Chicago, IL 1929

Kosslyn SM, Koenig O (1992) Wet mind, the new cognitive neuroscience. Free Press, New York

Lambert JH (1786) Theorie der Parallellinien. Magazin für die reine und angewandte Mathematik, 2–3:137–164, 325–358. Written about 1766; posthumously published by Bernoulli J

Lobachevsky NI (1829–1830, 1835–1838). Zwei geometrische Abhandlungen, aus dem Russischen bersetzt, mit Anmerkungen und mit einer Biographie des Verfassers von Friedrich Engel. B.G. Teubner, Leipzig

Lobachevsky NI (1891) Geometrical researches on the theory of parallels [1840]. University of Texas, Austin

Lobachevsky NI (1897) The "Introduction" to Lobachevsky's *New Elements of Geometry*. Trans Texas Acad 1–17. (Trans: Halsted GB). Originally published in Lobachevsky NI, Novye nachala geometrii, Uchonia sapiski Kasanskava Universiteta 3, 1835:3–48

Lobachevsky NI (1929) Pangeometry or a summary of geometry founded upon a general and rigorous theory of parallels. In: Smith DE (ed) A source book in mathematics. McGraw Hill, New York, pp 360–374
Magnani L (2001a) Abduction, reason, and science. Processes of discovery and explanation. Kluwer Academic/Plenum Publishers, New York
Magnani L (2001b) Philosophy and geometry. Theoretical and historical issues. Kluwer Academic Publisher, Dordrecht
Magnani L (2009) Abductive cognition. The epistemological and eco-cognitive dimensions of hypothetical reasoning. Springer, Heidelberg
Magnani L, Dossena R (2005) Perceiving the infinite and the infinitesimal world: unveiling and optical diagrams and the construction of mathematical concepts. Found Sci 10:7–23
Peirce CS (1931–1958) Collected papers of Charles Sanders Peirce. Harvard University Press, Cambridge. vol. 1–6, Hartshorne C, Weiss P (eds.), vol. 7–8, Burks AW (ed.)
Rosenfeld BA (1988) A history of non-euclidean geometry. Evolution of the concept of geometric space. Springer, Heidelberg
Saccheri G (1920) Euclides vindicatus. Euclid freed of every fleck. Open Court, Chicago. (Trans: Halsted GB). Originally published as Euclides ab omni naevo vindicatus, Ex Typographia Pauli Antonii Montani, Mediolani (Milan) 1733
Stroyan KD (2005) Uniform continuity and rates of growth of meromorphic functions. In: Luxemburg WJ, Robinson A (eds) Contributions to non-standard analysis. North-Holland, Amsterdam, pp 47–64
Tall D (2001) Natural and formal infinities. Educ Stud Math 48:199–238
Thagard P (2005) How does the brain form hypotheses? Towards a neurologically realistic computational model of explanation. In: Thagard P, Langley P, Magnani L, Shunn C (eds) Symposium "Generating explanatory hypotheses: mind, computer, brain, and world", Stresa, Italy. Cognitive Science Society, CD-Rom. Proceedings of the 27th international cognitive science conference
Thagard P (2007) Abductive inference: from philosophical analysis to neural mechanisms. In: Feeney A, Heit E (eds) Inductive reasoning: experimental, developmental, and computational approaches. Cambridge University Press, Cambridge, pp 226–247
Torretti R (1978) Philosophy of geometry from Riemann to Poincaré. Reidel, Dordrecht
Torretti R (2003) Review of Magnani L, Philosophy and geometry: theoretical and historical issues. Kluwer Academic Publishers, Dordrecht (2001). Studies in History and Philosophy of Modern Physics, 34b(1):158–160
Trafton JG, Trickett SB, Mintz FE (2005) Connecting internal and external representations: spatial transformations of scientific visualizations. Found Sci 10:89–106
Zhang J (1997) The nature of external representations in problem solving. Cogn Sci 21(2):179–217

Kant on the Generality of Model-Based Reasoning in Geometry

William Goodwin(✉)

Department of Philosophy, University of South Florida, Tampa, USA
wgoodwin@usf.edu

1 Kant's Relevance to Model-Based Reasoning

Kant can be seen as the philosophical ancestor of contemporary attempts to articulate the epistemological and practical significance of model-based reasoning. His recognition of the role of singular, immediate representations, or intuitions, in synthetic judgment–and thus in our ampliative claims about the world–is not only central to his philosophical program, but also grounded in many of the same insights that underwrite the contemporary MBR research program. Kant's account of the necessary role of intuition in synthetic judgment is introduced by way of his philosophy of mathematics. Mathematics, and most significantly Euclidean geometry, is what first opens Kant's eyes to the crucial role of what we would now call models in establishing contentful, novel claims about the world. Similarly, contemporary work in model-based reasoning in mathematics has also focused on Euclidean geometry (Giardino 2017). In this paper, I hope to bring out some of the ways that Kant's reflections on geometry not only anticipate this contemporary work, but also rest on the very same features of Euclidean proof. Furthermore, some of the challenges faced by Kant's model-based account of geometrical reasoning are still very much with us.

More specifically, I explore the relationship between two important aspects of Kant's philosophy of geometry. First is his claim that the truths of geometry are synthetic *a priori*, which I will spell out–in order to relate it to the MBR research program—as including the thought that geometry must be understood as model-based reasoning. Because models are more concrete or particular than the theories or linguistic representations that they allow us to reason about, appealing to models in geometrical reasoning creates a problem, which is explaining how such reasoning can be understood to support completely general conclusions. This is commonly known as the Generality Problem. After introducing these aspects of Kant's philosophy of geometry, I will contend that the very same facts about Euclidean reasoning that make Kant's doctrine of the syntheticity of geometry compelling serve to undermine the plausibility of his solution to the Generality Problem. Additionally, I try to motivate the thought that Kant's account fails at precisely the place where it should because the Euclidean practice that he is trying to account for lacked the resources to support a philosophically satisfying (by Kant's standards) account of the generality of it's results. Instead, following a suggestion from Ken Manders, the problem of generality brings out an often overlooked, but crucial, social and experimental aspect to model-based reasoning in Euclidean geometry.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 245–255, 2019.
https://doi.org/10.1007/978-3-030-32722-4_14

Given his central role in the model-based reasoning literature, a fitting way to introduce Kant's philosophy of geometry is by way of Peirce, who was—of course—a careful and insightful reader of Kant[1]. Unlike many interpreters of Kant that came both before and after him, Peirce, in my view, correctly recognized Kant as attempting to account for Euclidean reasoning (see also Friedman 1992). Not only does he correctly identify the features of Euclidean proof that motivated Kant's positive account of the synthetic *a priori* character of mathematics, but he also highlights Kant's strategy for explaining the generality of geometrical reasoning. Furthermore, he explicitly identifies both an observational and an experimental element within Kant's account of mathematics and then attempts to reconcile those with the necessity of mathematics. Peirce describes Kant's account as follows:

> Kant is entirely right in saying that, in drawing … consequences, the mathematician uses what, in geometry, is called a "construction", or in general a diagram, or visual array of characters or lines. Such a construction is formed according to a precept furnished by the hypothesis. Being formed, the construction is submitted to the scrutiny of observation, and new relations are discovered among its parts, not stated in the precept by which it was formed, and are found, by a little mental experimentation, to be such that they will always be present in such a construction. Thus, the necessary reasoning of mathematics is performed by means of observation and experiment, and its necessary character is due simply to the circumstance that the subject of this observation and experiment is a diagram of our own creation, the conditions of whose being we know all about. (Peirce, CP 3.560)

So to summarize Peirce's claims about Kant's account of geometry, he interprets Kant as recognizing and incorporating into his account the following observations:

(1) Diagrams play a crucial role in geometrical reasoning.
(2) Diagrams are constructed according to conditions furnished in the geometrical claim under consideration.
(3) Diagrams are subject to observation, and reveal novel features and relations not mention in the geometrical claim.
(4) Diagrams are subject to experimentation to insure the robustness of the observed novel relations.
(5) Observation and experimentation are compatible with the necessity of geometrical reasoning because geometers construct their diagrams according to known conditions.

2 Kant on Model-Based Reasoning in Geometry

Now, I will return to Kant, and try to briefly indicate the plausibility of Peirce's reading of him (for more detailed exposition, see Goodwin 2003). In particular, I want to emphasize the plausibility of finding both an observational and a manipulative/experimental element in Kant's account.

[1] The connection between Peirce, Kant, and model-based reasoning in geometry has been usefully developed in prior work by Magnani (2001), as well as extended, perhaps, to non-Euclidean geometry by Torretti (2001).

In his Inaugural Dissertation, Kant claimed:

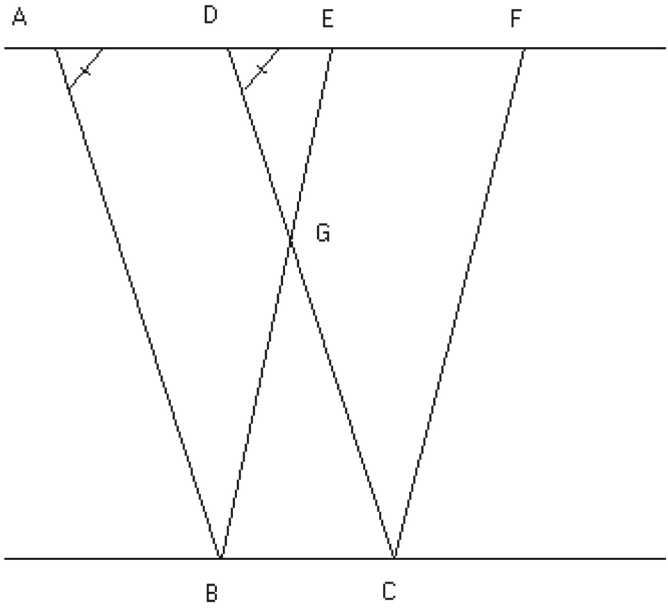
> Geometry employs principles which are not only indubitable and discursive, but which also fall under the gaze of the mind … geometry does not demonstrate its own universal propositions by thinking an object through a universal concept… It does so, rather, by placing it before the eyes by means of a singular intuition. (2:403; Kant 1992)

Throughout his writings, Kant supports his abstract characterizations of geometrical reasoning by describing features of the Euclidean practice. In particular, he draws attention to the figures constructed during the course of proof and the geometer's interaction with these figures. Furthermore, because intuitions, for Kant, are singular and immediate representations of objects, Pierce is right to associate Kant's claims about the role of intuitions in geometry, like those in the quote above, with the role of constructions and diagrams in Euclidean proof. A useful way to summarize what I call the positive aspect of Kant's doctrine of the syntheticity of geometry (see Goodwin 2003, 2010, 2018) is that the claims of geometry are underwritten by reasoning based on models of geometrical concepts constructed in pure intuition; further, the role of these models parallels the role of diagrams in Euclidean proofs. In his quote above, Peirce endorses the observations of Euclidean geometry upon which Kant's positive account is based, even if he neglects to mention the pure intuition that Kant invokes to in order to underwrite his philosophical account of this practice.

There are several features of the constructions-in-intuition that underwrite geometrical proof for Kant (and the diagrams in Euclidean proofs on which they are based) that make it appropriate to think of them as **models** of geometrical concepts. First, they are more specific and concrete than the concepts in accordance with which they are constructed. This comes out in both the fact that they can be observed to have properties or relations that are unexpressed at the conceptual level and the fact that they are manipulated during the course of proof in order to establish claims. Second, they are not instances of the concepts about which they facilitate reasoning. This comes out in the fact that some of them represent empty concepts (as in an reductio proof), which, though they have no instances, are nonetheless modeled during the course of geometrical proof (see Goodwin 2018). Additionally, Kant's constructions, and Euclidean figures, are intended to "express the concept, without impairing its universality" (A 713-4, B 741-2; Kant 1964), that is, they support general conclusions, not just statements about individuals. And finally, the diagrams that accompany Euclidean proofs are typically the product of auxiliary constructions that expand upon the figures set out in the theorem to facilitate the establishment of novel claims about them. It is because of all of these features that the constructions/figures underwrite the ampliative character of geometry reasoning. Thus Kant's positive account of the syntheticity of geometry, at least as I have interpreted it, rests on his recognition of Euclidean geometry as a form of model-based reasoning.

Ultimately, the plausibility of Kant's claims about geometry depend on his observations of how geometers go about proving things. To appreciate, then, the plausibility of the role that Kant attributes to constructions in geometry, it will be useful to look at an example of a Euclidean proof. The example I have chosen is a particularly simple theorem because it doesn't depend on any auxiliary constructions performed on the figure set out to represent the subject concept of the theorem. Nevertheless, the

proof does depend on both observation and experimentation/manipulation of the diagram. Euclid I.35 asserts (Heath 1956, 326–331) "Parallelograms which are on the same base and in the same parallels are equal to one another." The following diagram accompanies the proof:



The segment BC forms the base of the two parallelograms, BADC and BEFC. In modern terms, the goal of the proof is to show that these two parallelograms have the same area. The proof proceeds in two distinct stages. In the first stage, previous theorems are appealed to in order to establish that the triangles AEB and DFC are equal in area. By Proposition I.34, which has already been established, the opposite sides of parallelograms are of the same length. Thus BC is equal to AD, and BC is equal to EF, and then by transitivity AD is equal to EF. Likewise, by the same theorem, AB is equal to DC. Lastly, the angles BAD and CDF are equal by a property of parallel lines. Then, by the SAS criterion of triangle congruence, the triangles AEB and DFC are equal in area. In the second stage of the proof, the equality of the areas of the parallelograms is derived by combining the equality of the areas of the triangles from the first stage with topological features of the diagram. By subtracting the triangle DEG from both of these triangles (it is the area that they have in common in the diagram) one gets two quadrilaterals of equal area, ADGB and FCGE. By adding the same triangle, BGC, to each of these quadrilaterals, one gets that the desired parallelograms are equal in area.

Notice that the during the course of this proof, the geometer must observe that triangles exist in the diagram, even though triangles were never mentioned in the subject concept of the theorem. Instead these objects emerge during the course of the construction of the diagram. They are then proved to be congruent by what we now call the Side-Angle-Side Theorem. Finally, other regions that emerge during the initial

construction are subtracted from and added to the triangles to constitute the parallelograms mentioned in the theorem. The existence of the triangles that are crucial to the proof and the facts about the composition/decomposition of these emergent figures are only evident after construction of the diagram (see also, MacBeth 2010). Thus we can begin to appreciate why Pierce thought Kant was 'entirely right' to identify a role for models, and our observations of them, in geometrical reasoning.

Both the existence of new figures and the possible decompositions of emergent geometrical objects are things that the geometer comes to know by seeing them in the diagram, and so Kant, and Pierce after him, use the language of perception to describe the role of constructions in geometrical reasoning. We must put models of our geometrical concepts before the 'eye of the mind' to enable the observation such features, which are required in geometrical reasoning. Peirce also uses the language of experimentation to talk about how geometers must interact with their constructions during the course of geometrical proof. Kant likewise thinks of geometrical constructions–in-intuition as things that must be manipulated as part of geometrical proof. Ultimately, I think there are multiple features of geometrical proof that might be described as requiring manipulation or experimentation with the diagram, but in order to substantiate Peirce's reading, I want to bring out one case where Kant clearly recognizes the role of such manipulations.

As mentioned above, the proof of Euclid I.35 depended on establishing the congruence of triangles. All congruence theorems ultimately depend on Euclid I.4, which is what we now call the Side-Angle-Side Theorem. Euclid's theorem employs a proof technique that was used only three times in the *Elements*, and which has been eliminated in modern presentation of geometry (by adding additional axioms). This technique is called a Proof by Superposition, and in it the geometer considers what would happen if one figure was placed on top of another. Kant explicitly recognizes the importance of this proof technique, and appeals to it in support of his claim that geometry is synthetic. He says:

> In order to add something by way of an illustration and confirmation [of the fact that mathematics must construct its concepts in pure intuition], we need only watch the ordinary and unavoidably necessary procedure of geometers. All proofs of the complete congruence of two given figures (where the one can in every respect be substituted for the other) ultimately come down to the fact that they can be made to coincide. This is evidently nothing but a synthetic proposition resting upon immediate intuition. (4:284; Kant 1977)

Kant thus regards the manipulability of geometrical constructions (the fact that they can in some cases be "made to coincide") as another reason that they are required for geometrical proof. Thus, I hope to have now established why it was plausible for Kant to hold that geometers can demonstrate ampliative universal propositions only by recourse to models. These models are augmented by auxiliary constructions, manipulated, and observed to reveal novel objects, properties, and relations that are crucial to Euclidean proof.

3 On the Generality of Model-Based Reasoning in Geometry

Understanding geometrical truths to depend on model-based reasoning does, however, bring up some concerns about how to understand the certainty and generality of geometrical theorems. Kant puts the problem as follows:

> Mathematical knowledge is the knowledge gained by reason from the construction of concepts. To construct a concept means to exhibit *a priori* the intuition which corresponds to the concept. For the construction of a concept we therefore need a non-empirical intuition. The latter must, as intuition, be a single object, and yet none the less, as the construction of a concept (a universal representation), it must in its representation express universal validity for all possible intuitions which fall under the same concept. Thus I construct a triangle by representing the object which corresponds to this concept either by imagination alone, in pure intuition, or in accordance therewith also on paper, in empirical intuition …The single figure which we draw is empirical, and yet it serves to express the concept, without impairing its universality. (A 713-4, B 741-2; Kant 1964)

The concrete models used in geometrical proofs have all sorts of properties that are not shared by all "intuitions which fall under the same concept." For instance, any particular triangle is either equilateral, isosceles, or scalene, and so is, in that regard, not representative of all triangles. Nonetheless, as Kant notes in this quote, these models must "express universal validity" if they are taken to underwrite completely general geometrical claim. For instance, the particular triangle produced during the course of a Euclidean proof must support claims about all triangles. The Generality Problem is to explain how the geometer can establish general conclusions on the basis of–what Kant and Pierce to regard as essential–interactions with idiosyncratic individuals.

Proclus has a famous commentary on Euclid's Elements in which he addresses the Generality Problem in a way that usefully foreshadows Kant's account. Proclus says:

> For when they [mathematicians] have shown something to be true of the given figure, they infer that it is true in general, going from the particular to the universal conclusion. Because they do not make use of the particular qualities of the subjects but draw the angle or the straight line in order to place what is given before our eyes, they consider that what they infer about the given angle or the straight line can be identically asserted for every similar case. They pass therefore to the universal conclusion in order that we may not suppose that the result is confined to the particular instance. This procedure is justified, since for the demonstration they use the objects set out in the diagram not as these particular figures, but as figures resembling others of the same sort. (Morrow 1970, p. 162)

Here Proclus acknowledges that geometers make use of representations of idiosyncratic individuals during the course of proof; however, since those proofs don't make use of any of the idiosyncratic features of those individuals, but only those features that they share with all other instances of the subject concept, the proof applies equally well to those other individuals as well. We might say that the individual Euclidean proof can be interpreted as a sample of how to fill in a proof schema, which would apply to any individual that falls under the subject concept of the theorem. This approach to Generality is different from the one that we typically take in modern logic. It is what we might call an It's-how-you-don't-use-it solution to the Generality problem. Modern logic typically uses what Ken Manders has called a "representation-enforced unresponsiveness" strategy for establishing generality (Manders 2008a).

The idea of this strategy is to insist that the individuals we reason about are represented as having only those properties that are shared by the entire extension they are intended to stand for (see also Goodwin 2003 for an articulation of the different strategies). For Proclus' It's-how-you-don't-use-it strategy, instead of restricting the **representation**, it is the **reasoner** who is restricted. The reasoner must exercise discipline, using only those features of the individual representation that are shared with all individuals in the subject concept. According to Manders (2008a), and supported by other accounts of Euclidean reasoning (MacBeth 2010 and Netz 1999), this was the approach to generality characteristic of Euclidean reasoning, and perhaps ancient geometry more generally. However, Proclus' strategy naturally leads to the question of how the reasoner can know (with the certainty appropriate for a necessary truth, in Kant's case) that all the individuals falling under the subject concept share all of the features of the idiosyncratic individual appealed to in the proof?

Kant's discussions of the Generality problem all invoke the "conditions for the construction" of the intuitions that fall under the concepts geometers are reasoning about. You can see this in the following quote, where Kant explains that geometers don't make use of certain features of their constructions:

> I construct a triangle by exhibiting an object corresponding to this concept… The individual drawn figure is empirical, and nevertheless serves to express the concept without damage to its universality, for in the case of this empirical intuition we have taken account only of the action of constructing the concept, to which many determinations, e.g., those of the magnitude of the sides and angles, are entirely indifferent, and thus we have abstracted from these differences, which do not alter the concept of the triangle. (A 713-4, B 741-2; Kant 1964)

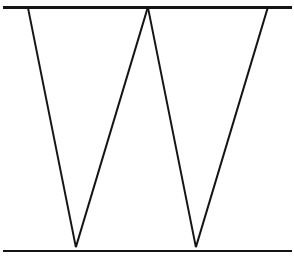
Just as for Proclus, generality for Kant is rooted in what features of their individual representations geometers don't use. Only features of individuals that are common to all constructions producible by a particular construction rule—or that, as Kant sometimes says, "follow from" the construction– are used in geometrical reasoning, and this is why such reasoning supports general conclusions. The same proof works for any individual constructed by those same construction rules because all that would be altered by substituting one product of these rules for another are "determinations" that do not figure into the original concept represented. Notice that this is where Peirce invoked a role for experimentation on geometrical figures. He claimed that "a little mental experimentation" was used to insure that the features observed in the individual diagram would be present in any figure constructed by the same rule. Furthermore, Pierce suggests that the necessity and certainty of geometrical reasoning that resorts to such experimentation is bound up with the fact that geometers make their own constructions and thus are aware of the rules by which they are produced. And indeed, this would seem to be a precondition for assessing whether all individuals produced by a certain rule have a property in common – you would have to know the rule by which those individuals were constructed. It is not as clear that knowing the rule is sufficient for knowing that all figures produced according to it would have all features appealed to during the course of proof, however.

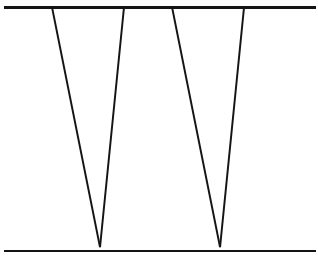
4 Contemporary Perspectives on Kant's Approach to Generality

Contemporary work on the role of diagrams in traditional geometry has born out, to a certain extent, the Its-how-you-don't-use-it explanation of the generality of Euclidean reasoning. Ken Manders has distinguished what he calls the exact and co-exact features of geometrical diagrams and used them to help explain the generality of traditional geometry (Manders 2008b, also Goodwin 2003). "Co-exact" conditions include part-whole relations, and other aspects of basic topology (such as the features observed in diagram during the proof of Euclid I.35). These features are stable across a range of distortions of the diagram; in other words, for at least a broad range of diagrams produced to meet the conditions set out in the theorem, these conditions continue to apply. "Exact" conditions such as lengths or measures, equalities and proportionalities, don't typically survive diagram variation. It is only, Manders observes, the co-exact features of diagrams that Euclidean proofs require the geometer to observe in the diagram. Because the sorts of features that geometers are called upon to observe in diagram are limited to co-exact features, and these co-exact features are stable across distortion of the diagram, the constructions and proofs of traditional geometry can be taken to apply across the distortion range of the co-exact features attributed during the proof. Insofar as the distortion range of the co-exact features attributed in the proof includes the extension of the subject concept of the theorem, then the proof exhibited using a particular diagram would establish that the same construction and reasoning would be applicable to any individual in that extension, and thus be fully general. So from this point of view Proclus and Kant are right: geometers must observe features of the diagrams and make use of them in their proofs, and generality results, in part, from limiting diagram-based attribution to co-exact properties of the diagram. This explanation of generality depends upon the geometer knowing (with certainty, if it is to meet traditional philosophical conceptions of mathematical knowledge) both what features are appealed to during the proof and what their distortion range is. Apart from Peirce's suggestion about experimentation, however, there is no explanation on offer of how the geometer could know these things.

Unlike Peirce and Kant, and because of the work of mathematicians like Proclus, Manders traces confidence in the generality of traditional geometrical proofs not simply to the geometer knowing the construction procedures for the diagrams he or she considers (and perhaps to their individual experimentation on these diagrams), but instead to the commentary tradition through which proofs are criticized and perfected over time. Manders has pointed out that traditional geometry used probing to ensure the stability of the co-exact features attributed during proof (Manders 2008b). That is, sustained experimentation on the proof and the stability of its attributions was an important part of the geometrical practice over time. The certainty and necessity that most philosophers have found in geometry, and that Kant was hoping to explain, might well be a result of the fact we moderns encountered Euclidean proofs only after thousands of years of successful probing, and not because each individual geometer has some necessary and certain access to the generality of their reasoning.

Apart from the issue of how the geometer can establish the distortion range of the features that are attributed in a geometrical proof, there is an additional issue facing the It-is-how-you-don't-use-it account of generality. This issue does not seem to have been acknowledged by Kant, Berkeley, Peirce, or even many contemporary philosophers trying to understand generality in Euclidean proof (MacBeth 2010, Netz 1999). The issue arises from a common feature of Euclidean proof, about which anyone who has worked through the Elements is aware, namely, proofs-by-cases. Proofs-by-cases show that the sorts of features attributed on the basis of the diagram are not standardly stable across the entire domain of the subject concept. That is, in such proofs, the fundamental assumption of the It-is-how-you-don't-use-it account of generality is false. Returning to Euclid I.35, it is easy to see that there are at least three distinct topological arrangements that are consistent with 'parallelograms with the same base and between the same parallels.' These topological arrangements differ in how the parallelograms are decomposed into other regions. The two arrangements besides the one presented in Euclid's proof are presented below:





Because Euclidean proofs often depend upon such topological decompositions attributed on the basis of the diagram (as we saw in our earlier proof), it will not generally be the case that the same proof will work for any instance of the subject concept. To relate this back to the its-how-you-don't-use-it account of generality, then, it is not the case that geometers only appeal to features attributable on the basis of the diagram in their proofs when those features are shared by all instances of the subject concept. Appeals to topological decompositions, which are often crucial in Euclidean proof, are appeals to features not always shared by the entire extension of the subject concept and so are not consistent with the proposed account of Generality. This problem is only magnified by the fact that further case splitting often occurs when auxiliary constructions are applied to the initial construction. In Euclid I.35, the original proof depended on these distinct decompositions, and so Proclus, as was typical in his commentary, describes the cases that Euclid left out. Again, it is the commentary tradition that fills in the cases whose consideration would be necessary in order to support a completely general conclusion.

Case-branching in Euclidean geometry arises from the fact that the linguistic characterizations of situations in the theorems of geometry do not capture all of the topological differences which might be necessary for geometrical proof. As a result, to support a general theorem, geometers would need to provide arguments (which may or may not be importantly distinct) for each of the topologically distinct cases. However, as Manders remarks:

> But how are participants to see whether a given selection of variant diagrams (after a construction complicates the diagram, or even at the outset of argument) is exhaustive of the possibilities which require separate argument? Traditional practice lacks procedure either to certify completeness of case distinctions, or to generate variants. (Manders 2008b, 107).

As a result, and contrary to modern formulations of geometry, individual traditional geometers were not in a position to prove the generality of their results. One might point out the known cases, and explain why you think you have them all, but generally one was not in a position to prove this. Instead, at least according to Manders, the commentary tradition in traditional geometry maintained a role for probing, which is where a proof is challenged as to the representativeness of its diagrams and the exhaustiveness of its case distinctions. Case-branch probing is thus another sense in which the traditional geometer depended upon experimentation (this time by others as well) with the diagram. So long as no objections were left unanswered, and no missing cases were identified, the geometer could be confident in the generality of results, but as Manders summarizes: "[C]ase distinction management remains in principle open-ended," the probing, "required in traditional practice is not supported by any clear cut or complete procedure, and therefore leaves geometric inference open-ended in a way which we moderns … hardly expect" (Manders 2008b, 120).

5 Conclusion

The role of diagrams in facilitating geometrical reasoning explains why Euclidean reasoning cannot be carried out at a purely conceptual level (as Kant understood that term) and instead must appeal to models of geometrical concepts. It also shows that it is not possible (in general) to understand these models as sharing all relevant features with every instance of the concepts under which they fall. Since Kant's solution to the Generality problem depends on being able to understand the individuals appealed to in geometrical reasoning in this way, his account of generality is not ultimately successful.

Proofs by cases arise in Euclidean geometry because of the mismatch between the conceptual and/or linguistic resources for characterizing geometrical situations and the features of those situations that are relevant to geometrical proofs. Diagrams bridge this expressive gap in Euclidean proof, and this is the source of the plausibility of Kant's claim that geometrical judgments are synthetic—that they require the consideration of individuals. At the same time, however, this mismatch insures that there will often be linguistic and/or conceptual characterizations of geometrical situations that are ambiguous over geometrical situations that differ in characteristics important to geometrical proof. This is the source of proofs by cases, and thus of some of the major difficulties for Kant's account of the generality of geometrical theorems. Thus it is the same features of Euclidean Proof that lend plausibility to Kant's claims about syntheticity which simultaneously undermine his explanations of the generality of geometry.

Kant did not have a good explanation of how one knows that all features appealed to in a diagram are shared with all other constructions which might fall under a

geometrical subject concept. Nor did Kant, to my knowledge, even address the challenges to generality posed by case distinctions, though they seem fatal to his approach. It is important to recognize, though, that the places where Kant's account broke down where places that the traditional practice also lacked resources to prove its assumptions, and instead depended on probing. This led others, who recognized the basic correctness of Kant's account of geometry, such as Pierce and Bolzano, to begin to develop the logical resources to support the generality of geometrical reasoning (see Goodwin 2010). With the development of modern logical approaches to generality it became possible to provide for the certainty and necessity of geometrical results while avoiding the traditional need for experimentation on the diagram.

References

Friedman M (1992) Kant and the exact sciences. Harvard University Press, Cambridge

Giardino V (2017) Diagrammatic reasoning in mathematics. In: Magnani L, Bertolotti T (ed) Springer handbook of model-based science. Springer

Goodwin W (2018) Conflicting conceptions of construction in Kant's philosophy of geometry. Perspect Sci 26(1):97–118

Goodwin W (2010) Coffa's Kant and the evolution of accounts of mathematical necessity. Synthese 172:361–379

Goodwin W (2003) Kant's philosophy of geometry, Ph.D. dissertation, University of California, Berkeley

Heath T (1956) The thirteen books of Euclid's elements, vol 1. Dover Publications, New York

Kant I (1964) The critique of pure reason (trans: Kemp-Smith N). St. Martin's Press, New York

Kant I (1977) Prolegomena to any future metaphysics (trans: Carus P, revised: Ellington J). Hackett Publishing Company, Indianapolis

Kant I (1992) On the form and principles of the sensible and intelligible world. In: Walford D (ed) Theoretical philosophy, 1755–1770 (trans: Walford D). Cambridge University Press, Cambridge

MacBeth D (2010) Diagrammatic reasoning in Euclid's elements. In: Van Kerkhove B, De Vuyst J, Van Bendegem JP (eds) Philosophical perspectives on mathematical practice, vol 12. College Publications, London

Magnani L (2001) Philosophy and Geometry: theoretical and historical issues. Kluwer Academic Publishers, Dordrecht

Manders K (2008a) Diagram-based geometric practice. In: Mancosu P (ed) The philosophy of mathematical practice. Oxford University Press, Oxford

Manders K (2008b) The Euclidean diagram (1995). In: Mancosu P (ed) The philosophy of mathematical practice. Oxford University Press, Oxford

Morrow G (1970) Proclus a commentary on the first book of Euclid's elements (trans: 1970). Princeton University Press, Princeton

Netz R (1999) The shaping of deduction in greek mathematics: a study in cognitive history. Cambridge University Press, Cambridge

Peirce C S (1958–1966) The Collected Papers: Peirce CS. In: Hartshorne C, Weiss P (ed). Harvard University Press, Cambridge

Torretti R (2001) Philosophy and geometry: theoretical and historical issues (Review: Magnani L). Stud Hist Philos Mod Phys 34:158–160



The Logic of Picturing: Wittgenstein, Sellars and Peirce's EG-beta

Rocco Gangle[1], Gianluca Caterina[1(✉)], and Fernando Tohmé[2]

[1] Endicott College, Beverly, USA
gcaterin@endicott.edu
[2] Universidad Nacional del Sur/CONICET, Bahía Blanca, Argentina

> The picture is a model of reality.
>
> L. Wittgenstein, *Tractatus Logico-Philosophicus 2.12*

Abstract. The semantics of picturing, broadly understood as an isomorphism between relevant relations among parts of a picture and relations constituting a state of affairs in some target domain, are a core feature of Wittgenstein's Tractarian theory of representation. This theory was subsequently developed by Wilfrid Sellars into a rich theory of language and cognition. In this paper we show that by recasting the positive fragment (without negation) of C.S. Peirce's beta level of Existential Graphs as a category of presheaves, the iconic coordination of syntax and semantics in the Wittgensteinian-Sellarsian picturing-relation may be represented formally in terms of the natural transformations in this category.

1 Introduction

A picture represents what it pictures by instantiating certain pictorial elements in certain relations similar in relevant respects to those of its object. There is thus a sort of structural mapping from the picture to the pictured. Such mappings may take a wide variety of forms and may be subject to various systems of representational convention. European painting before and after the development of linear perspective techniques in the fifteenth century, for instance, must be produced and viewed according to quite different rules of making and seeing. Taken more abstractly, this pictorial logic of structural mapping applies in a great many situations, including for instance musical or linguistic contexts that are in no way visual and thus do not literally involve "pictures". Yet the metaphor seems apt: it appears natural to represent structured objects or systems by other objects or systems that instantiate and thus share the relevant structure in a roughly pictorial fashion. From an epistemic point of view, knowledge and understanding of the represented object follows from appropriate investigation of properties of the representing picture (with "picture" taken here at times in its literal and at times its metaphorical sense). Truths about the represented object are "seen"

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 256–273, 2019.
https://doi.org/10.1007/978-3-030-32722-4_15

in, and thereby known through, the picture. Picturing thus becomes both a highly general mode of representation and potentially an epistemic methodology in various domains. For instance, the mathematics of category theory may be understood as the organization of abstract mathematical materials along such lines.

One strand of the philosophy of language has interpreted linguistic representation according to this, broadly speaking, pictorial logic. In particular, Ludwig Wittgenstein lays out such a view in his *Tractatus Logico-Philosophicus* and this understanding is picked up and carried forward in the second half of the twentieth century in the philosophy of Wilfrid Sellars. Distinct from although somewhat related to this way of thinking about language is work in formal logic by multiple thinkers, but in particular by Charles S. Peirce, that aims to develop iconic logical notations, that is, formal notations for logical properties and relations that in some ways or to some extent instantiate or exhibit those properties and relations themselves. This paper aims, quite modestly, to sketch out a viable point of local synthesis - determinate and particular in scope, yet potentially applicable in a variety of other contexts - between these two broad research programs, the one in philosophy of language and the other in formal logic. Mathematical tools drawn from category theory will underlie this connection and make it possible. In short, we will use category theoretical mathematics to show how Peirce's system of Existential Graphs may be understood perspicuously to represent the "picturing" relation as conceived by Wittgenstein and Sellars and, moreover, to represent this relation precisely by instantiating it at the appropriate level of abstraction. More particularly, we will recast a fragment of Peirce's beta level of Existential Graphs (without negation) as a category of presheaves in which natural transformations correspond to the Wittgensteinian/Sellarsian representation-relation of "picturing".

The paper is organized as follows: Sect. 2 characterizes the notion of representation as picturing that Wittgenstein presents in the *Tractatus* and outlines how Sellars develops that notion in the service of his naturalist theory of language and cognition. A list of desirable features for any formal notation that would represent and model this picturing relation is provided. Section 3 summarizes Peirce's EG-beta logical notation and introduces a fragment of that system without the syntactical elements of cuts and thus without negation in its semantics. Section 4 reconstructs the positive fragment of Peirce's EG-beta in a category-theoretical framework. More specifically, positive EG-beta graphs are represented as presheaves over a suitable base category. The semantics of picturing are provided by natural transformations in the functor category thus produced. Section 5 shows how a symmetric monoidal category may be generated over the functor category from Sect. 4 that offers a straightforward formal means for representing the patching together of multiple graphs and gluing their lines of identity. Finally, Sect. 6 summarizes the results and anticipates further work that would introduce logical negation into this framework.

2 Representation as Picturing

This section recalls the Wittgensteinian account of representation as picturing from the Tractatus, connects that account to the later use of this notion by Sellars, and introduces several desirable features for any formal notation that would aim to capture this particular type of relation.

2.1 Wittgenstein on Picturing

The second of Wittgenstein's seven primary propositions in the *Tractatus* (together with its dozens of sub-propositions) is devoted to what he calls the "representing relation" (*die abbildende Beziehung*) of a "picture" (*das Bild*) to the reality that the picture models, that is, the fact or state of affairs (*die Tatsache*) that it represents or pictures (*stellt dar*). The ground of this representation relation is the "logical form of representation" (*die logische Form der Abbildung*) that is shared by both the picture and the fact that is represented by the picture. Wittgenstein characterizes this logical form of the picture/fact as composed of two distinct types of features: objects or elements on the one hand, and the definite combinations of those elements or objects on the other. As Wittgenstein elaborates in [18]:

> 2.13 To the objects correspond in the picture the elements of the picture.
> 2.131 The elements of the picture stand, in the picture, for the objects.
> 2.14 The picture consists in the fact that its elements are combined with one another in a definite way.
> 2.141 The picture is a fact.
> 2.15 That the elements of the picture are combined with one another in a definite way, represents that the things are so combined with one another. This connexion of the elements of the picture is called its structure, and the possibility of this structure is called the form of representation of the picture.
> 2.151 The form of representation is the possibility that the things are combined with one another as are the elements of the picture.
> 2.1511 Thus the picture is linked with reality; it reaches up to it.

In the third and fourth of the *Tractatus*'s seven primary propositions, Wittgenstein applies this conception of picturing as modeling to linguistic propositions. He emphasizes that the pictorial character of linguistic representation is to some extent occluded by the standard form of language itself. As he puts it:

> 3.143 That the propositional sign is a fact is concealed by the ordinary form of expression, written or printed.
> (For in the printed proposition, for example, the sign of a proposition does not appear essentially different from a word [. . .]).
> 3.1431 The essential nature of the propositional sign becomes very clear when we imagine it made up of spatial objects (such as tables, chairs, books) instead of written signs.

> The mutual spatial position of these things then expresses the sense of the proposition.
> 3.1432 We must not say, "The complex sign '*aRb*' says '*a* stands in relation *R* to *b*"'; but we must say, "*That* '*a*' stands in a certain relation to '*b*' says *that aRb*".

Language, on this view, represents states of affairs by modeling those states via corresponding relations among the component parts of statements. Thus, at 4.01 we find the claim "The proposition is a picture of reality. The proposition is a model of reality as we think it is." Wittgenstein recognizes that this way of conceiving linguistic signs is somewhat surprising and counterintuitive. Nevertheless, he insists that all linguistic expressions that purport to describe what is the case, whether in spoken or written form, are actually in the appropriate sense *pictures* of the states of affairs they represent. As Wittgenstein puts it at 4.011:

> At the first glance the proposition–say as it stands printed on paper–does not seem to be a picture of the reality of which it treats. But nor does the musical score appear at first sight to be a picture of a musical piece; nor does our phonetic spelling (letters) seem to be a picture of our spoken language. And yet these symbolisms prove to be pictures–even in the ordinary sense of the word–of what they represent.

What kind of relationship is Wittgenstein suggesting here? It is clearly some kind of structural mapping from a source to a target domain. Even if the relative simplicity of this mapping is challenged by the developments in Wittgenstein's later philosophy (see in particular the somewhat more complex discussion of picturing in ([19], pp. 193–205), the view propounded in the *Tractatus*, because of its relatively straightforward and unnuanced character, appears susceptible to a possible translation into a more formal representational environment. The mathematics of category theory appears uniquely suited to formalize this kind of relationship and to make this insight, so far as possible, precise. This is what we will attempt to do in Sect. 4 and following. But first we will briefly examine how Wittgenstein's core idea here is carried forward and elaborated further in the work of Wilfrid Sellars.

2.2 Sellars on Wittgenstein on Picturing

Wilfrid Sellars develops the Wittgensteinian idea of representation as picturing into a sophisticated naturalist theory of language and cognition, a research program aptly described in [11] as "naturalism with a normative turn". In chapter 5 of [16], among other places, Sellars focuses on the notion of picturing in the *Tractatus* and distinguishes carefully between the sort of complexity that can result from the juxtaposition and composition of different kinds of ("atomic") pictures and the deceptively similar mode of complexity and composition (which he calls "molecular") which is specifically *logical*, that is, the result of formal logical operators of various kinds. The former are empirical features of complex pictures. The latter are distinctively logical properties. As Sellars puts it ([16], p. 105):

> the mode of composition by virtue of which a number of atomic statements join to make a complex picture must not be confused with the mode of composition by virtue of which a number of atomic statements join to make a molecular statement. In other words, we must distinguish "pictorial" from "logical" complexity.

Sellars offers the example of a pictorial code using letters in which different fonts represent various properties and particular spatial arrangements of letters represent different relations. In this way, what might seem to be a kind of logical notation is, from Sellars's point of view, more appropriately understood as pictorial in character. Thus the notational picture

$$\mathbf{a}_{\mathrm{B}}$$

is understood by Sellars to represent a pair of objects named 'a' and 'b'. The given notational configuration might then in its own order signify (1) that the object named a has some property F (because the name is in bold font), (2) that the object named b has some property G (because its name is in upper case), and (3) that the object named a stands in some relation R to the object named b (because the latter name is positioned to the lower right of the former).[1]

Sellars uses this syntactical or notational difference to distinguish a broadly platonistic from a nominalistic approach to linguistic representation. Sellars points out that the compositionality of relations in a notation which instantiates relations directly provides a natural means for simultaneously representing multiple relations in a complex or conjoined system without requiring any additional representational machinery. In particular, one does not require additional symbols to represent logical relations such as conjunction or additional rules to establish at least some logical properties such as the commutativity or associativity of conjunction. Merely empirical features of the notation (such as the juxtaposition of written signs on a common sheet) are formally sufficient to model some of the features that will eventually be understood as abstract logical properties (such as the commutativity and associativity of conjunction). Intrinsic characteristics of pictures are thus already anticipations or precursors of certain formal logical properties.

In this respect Sellars emphasizes "the central role, in an adequate nominalistic theory of linguistic representation, of a stratum of complex representations (maps), the constituents of which have an *empirical* structure" ([15] p. 75). The "merely" empirical character of these components of complex representations is important for Sellars because it provides a potential basis for explaining how purely logical elements of cognition (such as the logical conjunction of properties or propositions) might emerge from, roughly, inductive generalizations of empirical features of practical language use. This explanatory strategy for moving from the empirical (and its ambient "space of causes") to the ideal (and its constituent "space of reasons") is at the core of Sellars's ambitious project of philosophical naturalism and is also at the heart of Robert Brandom's appropriation of Sellars

[1] This example is Sellars's own as presented and discussed in [16], p. 106.

for his broadly Hegelian program of philosophy's reflective "making explicit" of implicit norms of reasoning (see [2] and also [3] and [4]).

Sellars clarifies how the pictorial understanding of notation provides a means for finessing this passage from the empirical to the logical in [15] (p. 75–76):

> The formation rules of the language pick out items having certain *empirical* forms to have *logical* form in the sense of *predicational* form, i.e. they function as atomic sentences, but do not, *as functioning in this stratum*, have logical form in the sense of *undergoing* logical operations (truth-functional combination, quantification), although, as constituents of a *representational system*, they are *subject to* these operations either directly (by wearing, so to speak, another hat) or *indirectly* by being correlated with (translatable into) other designs which are directly subject to these operations.

To express Sellars's point in a Peircean idiom, we might say that a nominalist approach to linguistic representation (that is, one that rejects so far as is possible the introduction of any ineliminable references to universals) is greatly facilitated by an appropriately *iconic* form of notation. This is ultimately because linguistic representation itself is, at heart, iconic in nature. Realism (anti-nominalism) concerning universals may, from this point of view, be *explained* as a natural, albeit mistaken, inference - a kind of transcendental illusion - drawn from imperspicuous features of language itself, namely those identified by Wittgenstein at 3.143 in [18] as discussed above. There is thus reason to suspect (and to hope) that a good notation for picturing will aid philosophical reflection on language – if, that is, the approach of Wittgenstein and Sellars is, in broad strokes, correct.

2.3 Desiderata: Design Features for a Perspicuous Picturing Notation

Given the conception of representing as picturing developed by Wittgenstein and Sellars, we can identify several characteristics that would be desirable for any regimented notation that would make use of this conception.

1. The objects and relations in the notation ought to be strongly visually distinguishable as distinct *types* of compositional elements so as to avoid the sort of confusion identified by Wittgenstein at 3.143 in [18] and further analyzed by Sellars in the context of his nominalist interpretation of language.
2. The notation should support a general semantics of picturing that might eventually be extended to full first-order logic as seamlessly as possible. This would be in order to facilitate the conceptual and analytical transition from the empirical features of linguistic representation to the formal logical properties of the Sellarsian "space of reasons".
3. There ought to be formal notational provisions for cutting up pictures and patching them together with or without overlaps. This is after all how (literal) pictures actually work - a photograph can be torn in two to produce two pictures, for instance, and two photographs of the same scene might overlap on certain represented elements.

The following reconstruction of a fragment of an iconic logical notation developed by Peirce is meant to answer to these three desiderata.

3 $EG_{\beta+}$: Peirce's EG-Beta without Negation

From the mid-1880s up to his death in 1914, Charles S. Peirce developed a system of graphical logic which he came to call the Existential Graphs (EG). The system of EG consists of three levels: alpha, beta and gamma, corresponding to classical propositional logic, first-order logic with identity, and various modal logics, respectively. A general introduction and summary of EG may be found in [13]. What is of particular interest in the present context is that logical relations are represented in EG by a notation that instantiates topological analogues of those relations themselves. In other words, the notation of EG is *iconic* in a way that helps to guide reasoning processes of hypothesis-formation and testing (as analyzed, for instance, in [6] and [7]).

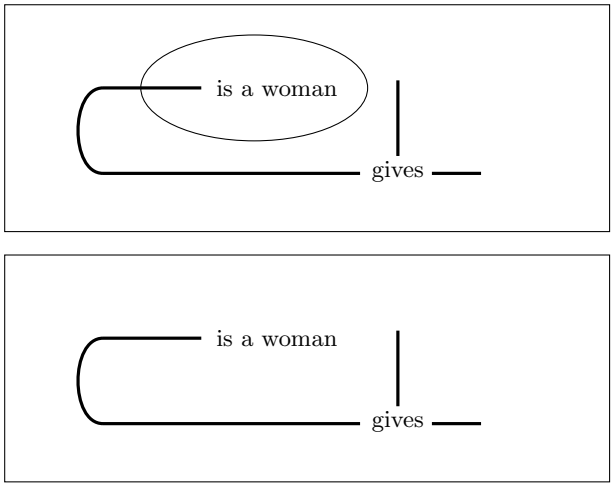
We claim that the beta level of Peirce's Existential Graphs as restricted to graphs without cuts meets the above criteria for a perspicuous picturing notation. We call this fragment of Peirce's beta graphs the *positive fragment* of EG-beta and label it $EG_{\beta+}$.

Detailed studies of Peirce's EG-beta may be found in [8,12,13,17] and [20]. We consider here only those EG-beta graphs that do not involve *cuts* (which Peirce also calls *seps*). Thus, the graphs with which we are concerned are composed only of *lines of identity* and *spots*, which correspond to existential quantifiers and n-ary relations as scribed on the *Sheet of Assertion*. Since cuts represent logical negation in Peirce's notation, the system $EG_{\beta+}$ lacks the capacity to represent negation.[2] This means that $EG_{\beta+}$ is reduced significantly in its expressive power, but it is also tremendously simplified. The resulting system is quite similar in spirit to *regular logic*, which has received increasing interest in recent years due its surprisingly varied applicability (see [5] and [9]). A pair of examples should suffice to show informally how the simplified system works.

Of the pair of EG-beta graphs pictured below, the top graph encloses the property (i.e. the unary relation) "is a woman" together with its single hook within a cut. The same line of identity is connected to the first hook of the triadic relation "gives". Thus, the graph may be understood to assert that *someone who is not a woman gives something to someone*. The second graph below represents the same situation but without the cut. It may be read as saying that *someone who is a woman gives something to someone*. Without cuts and, in particular, without nested cuts, universal quantification cannot be expressed in the reduced system and the truth-preserving transformation rules proposed by Peirce are essentially rendered superfluous.[3]

[2] The reader might compare the variant of Peirce's system detailed in [1] in which the cut as classical negation is also absent but is there replaced with a new type of cut with a non-classical interpretation.

[3] Because of our focus on the pictorial character of the graphs and not their strictly logical properties, we do not address the transformation rules of Peirce's system in this paper.



We intend to maintain the Sellarsian standpoint according to which relational predicates in the notation are understood iconically. In other words, they do not abstractly represent universals but are rather concrete instantiations (*empirical* instantiations) of relational counterparts of the very relations at issue. On this view, the "spots" in Peirce's beta-graphs do not merely *stand for* particular relations but are rather understood as doing so first and foremost by instantiating determinate relations themselves.

The point is a somewhat subtle one, but it is important. It is, in fact, crucial to Sellars's approach as derived from Wittgenstein. In the above example, for instance, the written word "gives" in the EG-beta graph links three ordered lines of identity. This syntactical triadic relation in the graph, from the Wittgensteinian/Sellarsian point of view, is capable of supporting its representational function *of* the worldly triadic relation of giving (someone gives something to someone) because it itself *is* a triadic relation (three lines of identity as connected to the left, top and right sides, respectively, of the sign-token "gives"). This is the core idea of the iconic "picturing" function of the graphs that will be formalized in what follows via category theory.

4 A Generic Figures Reconstruction of $EG_{\beta+}$

We now cast the positive fragment of Peirce's EG-beta in a category-theoretical setting. This approach allows us to generate all well-formed $EG_{\beta+}$ graphs in a single step by characterizing them as functors from an appropriately structured base category to the category **FinSet** of finite sets and functions.

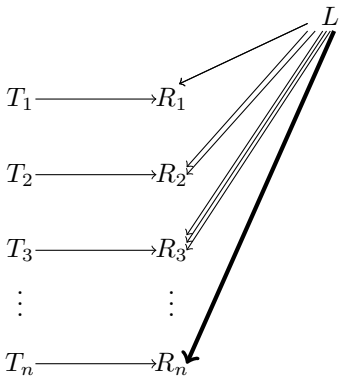
This characterization of $EG_{\beta+}$ graphs as presheaves engages the graphs at a useful level of abstraction. Isomorphic presheaves characterize graphs that are identical in structure in a sense that respects the core insight of Wittgenstein and Sellars but lifts this insight to a higher level of abstraction. In this way the functor category captures the structural features of the picturing relation itself, picturing it at a second-order level of abstraction.

For this reason, the traditional conception of a sharp distinction between syntax and semantics is out of place in the context of the Wittgensteinian-Sellarsian notion of picturing. Part of Wittgenstein's original motivation for understanding representation in these terms was to preserve a unified approach to the "pictures" and what is thereby "pictured". Sellars develops this unified approach in his own distinctive naturalist approach to language and cognition. Yet this puts the logician in something of a bind. It would seem that two incompatible requirements are presented. On the one hand, syntax and semantics must be elements of the same system. On the other hand, the possibility of lifting this relative symmetry to the asymmetrical relationship of logical form as applied to concrete models should be preserved. The categorical approach to Peirce's graphs provides a means for resolving this apparent incompatibility.

Because of this non-standard approach to the syntax-semantics difference, the titles of the sections below should be taken with a grain or two of salt. What is intended is simply to present the analogues within the present approach of the concepts of syntax and semantics as these latter have been more or less regimented and standardized in modern logic.

4.1 Syntax

The iconic syntax of Peirce's EG_{β^+} is given by the class of contravariant functors from the category pictured below, which we notate $\mathcal{B}$, into the category **FinSet** of finite sets and functions between these.



More precisely, the category $\mathcal{B}$ consists of objects and morphisms specified as follows:

- Objects: $\{T_i\}_{i\in\mathbb{N}}, \{R_i\}_{i\in\mathbb{N}}, L$
- Morphisms:
  - identities;
  - a collection of morphisms $\{t_i\}_{i\in\mathbb{N}}$ where

$$t_i : T_i \longrightarrow R_i$$

– for each $i \in \mathbb{N}$ a collection of morphisms $\{r_i^j\}_{j=0,1,\ldots,i}$ where

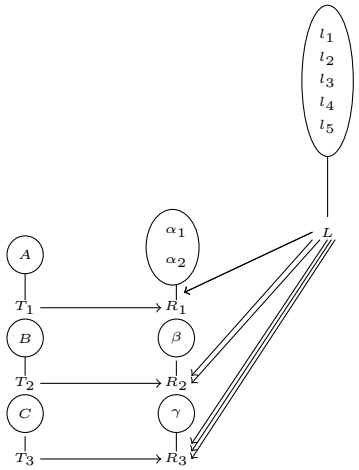
$$r_i^j : L \longrightarrow R_i$$

Formally, then, a EG_{β^+} graph is a functor $G : \mathcal{B}^{op} \longrightarrow \mathbf{FinSet}$ such that there is some n such that for all $m > n$ $G(T_m) = \emptyset$. This latter condition simply ensures for the sake of tidiness that every graph has a maximal relation arity.

We convert any such functor into Peirce's notation in the following way:

- Each element of $G(L)$ is drawn on the Sheet of Assertion as a line of identity.
- Each $G(T_n)$ is assigned a distinct type of relation-sign with n hooks which may be ordered according to whatever method is convenient (here, we understand them to be ordered clockwise starting from the left).
- Each $G(R_n)$ is drawn on the Sheet of Assertion as a token of relation-sign-type $G(t_n)[G(R_n)]$ [4].
- The n hooks of each relation-sign-token are attached to lines of identity according to the functions $G(r_n^j)[G(R_n)]$. Specifically, hook j of relation-sign-token $G(R_n)$ is attached to line of identity $G(r_n^j)[G(R_n)]$.

An example should make this clear. Consider the diagram below, which represents a functor from $\mathcal{B}$ into **FinSet**.



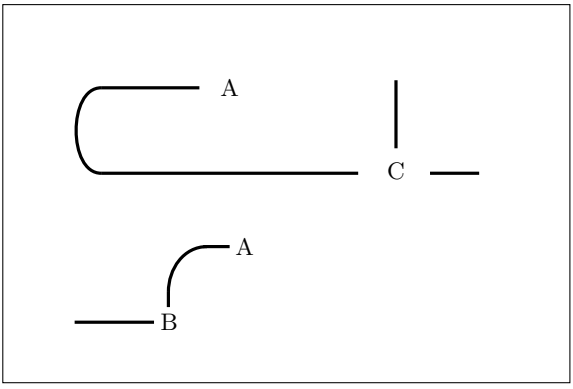
The ovals above each of the category objects represent the sets to which the functor G sends those objects. All objects that are not shown, such as T_4 are understood to be sent to the empty set. For instance, $G(R_1)$ is the two-element set containing α_1 and α_2. The three functions $G(t_1)$, $G(t_2)$ and $G(t_3)$ are completely determined since their codomains are singletons (the reader should keep in mind the contravariance of the functor).

[4] Here the notation $G(t_n)[G(R_n)]$ represents the action of the "lifted" function t_n via the functor G as it acts on the "lifted" set R_n via the same functor G.

We may stipulate that the remaining functions are defined as follows (the functions are listed in the top row and their argument in the leftmost column):

	$G(r_1^1)$	$G(r_2^1)$	$G(r_2^2)$	$G(r_3^1)$	$G(r_3^2)$	$G(r_3^3)$
α_1	l_1					
α_2	l_5					
β		l_4	l_5			
γ				l_1	l_2	l_3

The resulting beta graph may then be pictured as below:



The category $\mathbf{FinSet}^{\mathcal{B}^{\mathrm{op}}}$ consists of all such functors as objects, with natural transformations between functors as morphisms.

In general, given two categories $\mathcal{C}$ and $\mathcal{D}$, the functor category $\mathcal{D}^{\mathcal{C}}$ is defined as follows:

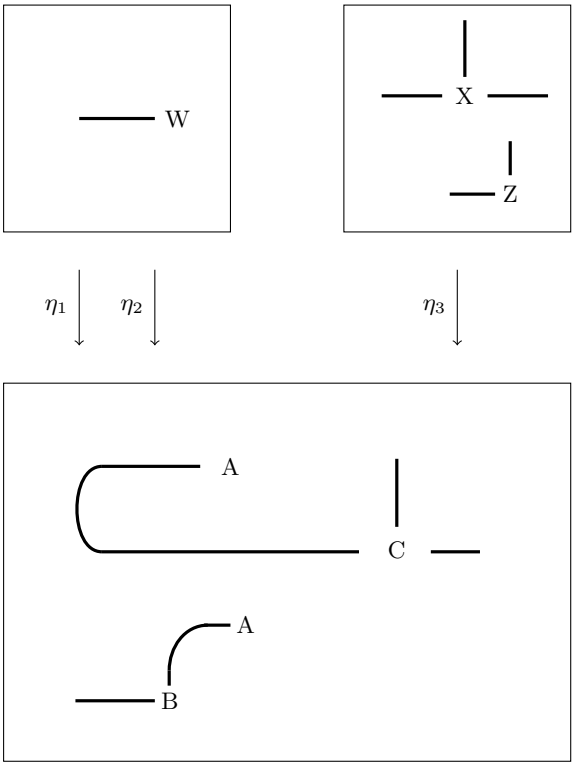
- The objects of $\mathcal{D}^{\mathcal{C}}$ are all *functors* $\mathcal{C} \to \mathcal{D}$.
- The arrows of $\mathcal{D}^{\mathcal{C}}$ are all *natural transformations* between functors $\mathcal{C} \to \mathcal{D}$.

Natural transformations are thus morphisms between functors: given two functors $F \in \mathcal{D}^{\mathcal{C}}$ and $G \in \mathcal{D}^{\mathcal{C}}$, a natural transformation between F and G is a family of morphisms η_O parametrized by the objects $O \in \mathcal{C}$ such that the following diagram commutes for any two objects A and B that are connected by a morphism f in $\mathcal{C}$:

$$\begin{array}{ccc} F(A) & \xrightarrow{F(f)} & F(B) \\ \Big\downarrow{\scriptstyle \eta_A} & & \Big\downarrow{\scriptstyle \eta_B} \\ G(A) & \xrightarrow{G(f)} & G(B) \end{array}$$

In this way $\mathbf{FinSet}^{\mathcal{B}^{\mathbf{op}}}$ names the category of contravariant functors from $\mathcal{B}$ to **FinSet** and natural transformations among these. For simplicity's sake, we rename the category $\mathbf{FinSet}^{\mathcal{B}^{\mathbf{op}}}$ as $\mathcal{EG}_{\beta^+}$.

Morphisms in $\mathcal{EG}_{\beta^+}$ represent structure-preserving maps from one functor into another that are directly reflected by maps from the beta graph representing the former into the graph representing the latter. Again, an example (a picture) will help to make this clear. We may take the beta graph constructed in the previous section as the codomain of three different natural transformations, as shown below. The two natural transformations on the left, labeled η_1 and η_2, have as domain the graph with a single line of identity and a single property W. The natural transformation on the right, η_3, has as domain the graph with a single triadic relation X and a single dyadic relation Z and five distinct lines of identity.



For both η_1 and η_2, the unary relation-type W in the natural transformation's domain is forced to be mapped to the relation-type A in the codomain. This is because A is the only unary relation-type available in the codomain, and natural transformations preserve the arities of relation-types. However, there are two distinct relation-tokens of A in the lower graph. There are thus two distinct ways the relation-token W may be mapped into the codomain, namely to each of the two different relation-tokens A. These two distinct maps thus exhaust the

possible natural transformations from the functor represented by the beta-graph on the upper left into the functor represented by the lower graph.

In a similar manner, there exists exactly one natural transformation, here labeled η_3, from the functor represented by the graph on the upper right into the functor represented by the lower graph. The triadic relation-type X is constrained to be mapped to the relation-type C and since there is only one token of type C, the token of X must be mapped there. By the same logic, the type and token Z must be mapped to the type and token B.

The reader should note that because the functors underlying the beta-graphs carry no more than the information necessary to identify the *structure* of the graphs – and not their contingent content – the "naming" of the relation-types by letters such as A or B is in fact arbitrary. In fact, the relation-types are specified only up to their identifiability, that is, only up to their identity with or difference from one another. Thus, each beta-graph as generated from a given functor is properly understood as a representative of an equivalence class of graphs with isomorphic structure.

In the following section, we show how morphisms in $\mathcal{EG}_{\beta+}$ (i.e. natural transformations between functors) correspond to the picturing relation itself as analyzed by Wittgenstein and Sellars.

4.2 Semantics

It should be clear that the notational interpretation of each object of $\mathcal{EG}_{\beta+}$ is somewhat arbitrary. Other ways of "picturing" the functor (the elements of the sets $G(L)$, $G(R_n)$ and $G(T_n)$ and the functions $G(t_n)$, $G(r_n^j)$, etc.) may be devised. Nonetheless, the idea might suggest itself to the reader that the notation described above is in certain key respects somewhat optimal, given the intended informal interpretation of the functors as pictures of objects standing in discernable relations to one another. The objects are here represented by featureless lines (in some sense the minimal representation of a thing that does not import additional qualities into the representation), and the relations are represented by tokens of types that are specified only up to their recognizability *as* notational types.

This need for a choice of how to instantiate the merely formal dimension of the syntax (the functor) in the material or empirical notation of Peirce's graphs might appear to suggest that the lifting of Peirce's graphs to the abstract level of category theory introduces an unnecessary complication. Why not, after all, just work with the formalism of the graphs themselves? Yet what seems to be a complication and the introduction of an arbitrary choice is in fact the key to using this formalism to represent the picturing "form of representation" itself.

The basic idea is that the base category carries the higher-order structure of how objects, relation-types and relation-tokens are structured with respect to one another. The constituent objects and arrows of the category may be understood roughly as intrinsic rules for construction, somewhat as symplectic geometrical objects may be conceived as rules for their own construction. A particular functor G then instantiates this higher-order, constructive structure in some determinate

way. The functor is relatively concrete with respect to the base category but still abstract in the sense that the sets and functions that determine the functor are not yet further interpreted. This next step of interpretation consists of assigning some "object" or "name" to each element of $G(L)$, some relation-type to each element of $G(T_n)$ and some actual relation of a certain type to each element of $G(R_n)$. The functions in the codomain of the functor provide the necessary connections ensuring which relations are of what type and which objects/names are related by what relations and exactly how. In short, the functions specify how the relations relate the objects/names.

So the interpretation given above that allows for the construction, given a functor G, of some particular Peircean beta-graph represents simply one way – one possible choice – of how to instantiate G concretely. Other choices will correspond to other ways to compose a "picture" or, equivalently, to organize or interpret a state of affairs.[5] This fact reflects the core idea shared by Wittgenstein and Sellars that pictures (representations) and states of affairs do not live in distinct ontological realms but are rather both constituents of one and the same world.

Thus, we may take the objects of our functor category to represent states of affairs (or, equivalently, pictures) in a quite general sense. Peirce's graphs become just one type of picture from this point of view, but a type that can be isomorphically substituted for *any* such picture or state of affairs via the underlying functors. Essentially, Peirce's graphs instantiate the formal relations characterized by the functors in a visual form that regularizes the difference between object-terms and relation-terms and makes this difference visually salient. Maps between these functors (natural transformations) serve as the morphisms in our category; they map objects to objects and relations to relations across functors in the appropriately structure-preserving way. Such natural transformation maps then, from this categorical point of view, just *are* the picturing relations at stake.

The upshot is that one and same formal framework, the functor category $\mathcal{EG}_{\beta^+}$, captures *both* the positive fragment of Peirce's EG-beta notation *and* the states of affairs themselves. What both sides of the representation relation share, namely the *structure* that is common to both and makes the picturing relation possible is represented at the appropriate level of abstraction by the functor itself.

[5] To organize pictures/states of affairs in a Tarski-style set-theoretical way for instance, given some functor G, one could assign the elements of $G(L)$ to a chosen set M, the elements of each T_n to a subset of M^n and each R_n to an element of M^n (that is, an n-tuple over M). The functions $G(r_n^i)$ would then be assigned to the obvious projection maps, and the functions $G(t_n)$ would be required to send n-tuples $< a_1, \ldots, a_n >$ to subsets S such that $< a_1, \ldots, a_n > \in S$.

5 Generating a Symmetric Monoidal Category on Peirce's EG-beta

Given two objects (i.e. two functors) F and G in $\mathcal{EG}_{\beta^+}$, the categorical *coproduct* of F and G, designated $F \oplus G$, is constructed as follows:

- for any object X of the base category $\mathcal{B}$

$$(F \oplus G)(X) = F(X) \sqcup G(X)$$

  where $A \sqcup B$ is the disjoint union of the two sets A and B.
- for any morphism $f : X \longrightarrow Y$ in $\mathcal{B}$

$$(F \oplus G)(f) = f \sqcup g$$

  where

$$f \sqcup g : F(X) \sqcup G(X) \longrightarrow F(Y) \sqcup G(Y)$$

  and

$$(f \sqcup g)|_{i_{F(X)}(F(X))} = F(f)$$
$$(f \sqcup g)|_{i_{G(X)}(G(X))} = G(f)$$

  where $i_{F(X)}$ and $i_{G(X)}$ are the canonical inclusions of $F(X)$ and $G(X)$ into $F(X) \sqcup G(X)$.

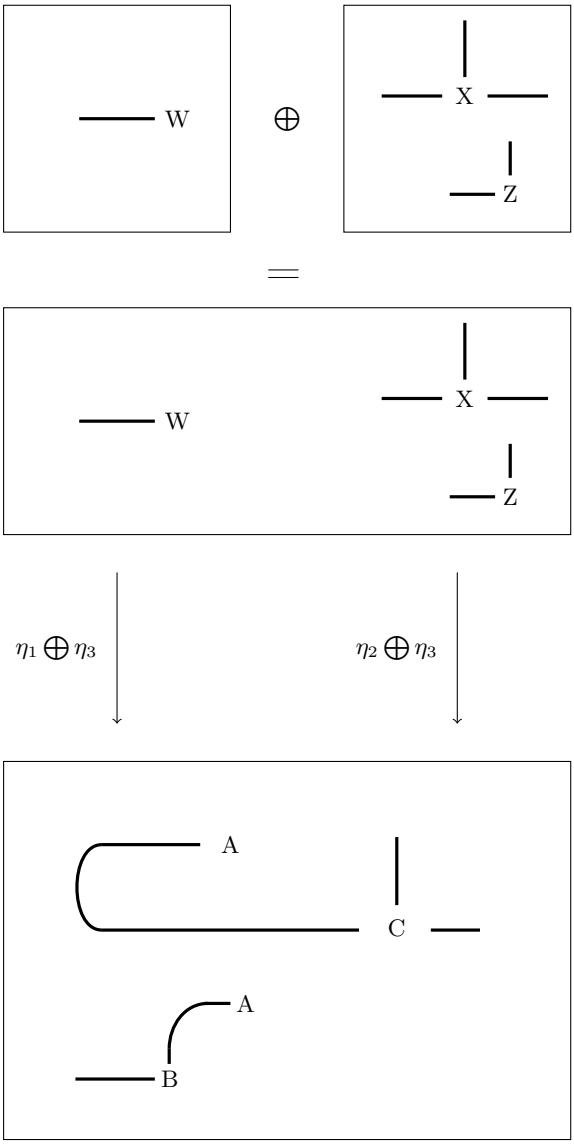
The coproduct defined set-theoretically in this way is the categorical coproduct of objects in $\mathcal{EG}_{\beta^+}$ as defined via the standard universal property. Such a coproduct exists for every pair of functors F and G in $\mathcal{EG}_{\beta^+}$. Every category with all finite coproducts induces a *symmetric monoidal category* (SMC) where the monoidal product is the category coproduct. For details, as well as the full definition of a symmetric monoidal category and a survey of some mathematical applications, see [10], pp. 184, 251–266.

In the present case, the monoidal product (here, the categorical coproduct) corresponds to the juxtaposition of EG beta-graphs on one and the same Sheet of Assertion. Specifically, given some collection of EG-beta graphs represented by functors, say $G_1, G_2, \ldots G_n$, the graph represented by the monoidal product $G_1 \oplus G_2 \oplus \cdots \oplus G_n$ will simply be the juxtaposition of all the graphs on a single Sheet of Assertion. The monoidal *unit* may then be understood as the blank Sheet of Assertion itself (the functor that takes all objects of $\mathcal{B}$ to the empty set $\emptyset$). Naturally enough, any graph G juxtaposed with a blank sheet just is G itself. Here, it should be clear that the visual (and hence syntactic) relation of juxtaposition corresponds to the logical (semantic) relation of conjunction.

As pointed out in [14], there are many kinds of monoidal categories with additional structure that have an associated graphical language of *string diagrams*. Further research could apply such methods to the gluing and embedding of Peirce's EG beta-graphs in the present context.

We then have the situation such that maps in the symmetric monoidal category correspond to embeddings of graphs, including overlaps and gluings of

lines of identity (which thus play a dual role at the semantic level as existential quantifiers and relations of equality). As a simple example, we will show how the previous example from Sect. 4.1 can be recast via the monoidal products of the two graphs in the domain of the natural transformations η_1, η_2, η_3. Notice that here one line of identity has been glued.



In this way, string diagrams in the SMC over $\mathcal{EG}_{\beta+}$ promise to provide a useful notation for tracking embeddings and gluings of multiple graphs into others. Since graphs themselves are being considered here as formal notations for the

relevant structure of pictures in the Wittgensteinian/Sellarsian sense, the SMC may be thought of as a logic for the controlled manipulation of the notation itself. Furthermore, since the notation itself, in accordance with the logic of picturing, shares the relevant structure with the states of affairs it represents, this controlled manipulation corresponds directly to possible overlaps, intersections and embeddings in the realities thereby pictured.

6 Conclusion

The results of the previous sections may be summarized as follows:

1. The fragment of Peirce's EG-beta graphical notation without cuts (thus without negation and universal quantification) was proposed as an iconic formal notation that would both instantiate and represent the picturing-relation of representation as analyzed by Wittgenstein and Sellars.
2. A formal reconstruction of the positive fragment of Peirce's beta-graphs was cast in a category-theoretical framework as the category $\mathcal{EG}_{\beta+}$ of contravariant functors from a base category $\mathcal{B}$ into **FinSet** and natural transformations between such functors.
3. Natural transformations in $\mathcal{EG}_{\beta+}$ were shown to represent the Wittgensteinian/Sellarsian picturing-relation as such.
4. Construction of a symmetric monoidal category on the basis of coproducts in $\mathcal{EG}_{\beta+}$ captures the dynamics of juxtaposition of graphs on a common sheet of assertion with maps corresponding to (pictorial/graphical) overlaps and gluings.

Formalizing the picturing relation in terms of Peirce's positive EG-beta graphs not only helps to clarify and make precise the insights of Wittgenstein and Sellars. It also promises to contribute to the Sellarsian program of understanding the passage from empirical to logical relations.

Given that the positive fragment of Peirce's EG-beta is already part of a notation for full first-order logic with identity and that it is also useful for capturing the logic of picturing, it seems promising to think that the logical operation of negation, together with all the logical features that come along with it, might be smoothly added to the framework provided above. If so, then the transition central to the program of Sellarsian naturalism from merely empirical features and regularities in practical language to fully-fledged logical properties among concepts and propositions would become significantly more tractable.

References

1. Bellucci F, Chiffi D, Pietarinen A-V (2018) Assertive graphs. J Appl Non-Classical Logics 28(1):72–91
2. Brandom R (1994) Making it explicit: reasoning, representing, and discursive commitment. Harvard, Cambridge

3. Brandom R (2015) From empiricism to expressivism: brandom reads sellars. Harvard, Cambridge
4. Brandom R (2009) Reason in philosophy: animating ideas. Harvard, Cambridge
5. Butz C (1998) Regular categories and regular logic. BRICS LS, Aarhus
6. Caterina G, Gangle R (2016) Iconicity and abduction. Springer, New York
7. Caterina G, Gangle R (2013) Iconicity and abduction: a categorical approach to creative hypothesis-formation in peirce's existential graphs. Logic J IGPL 21(6):1028–1043
8. Dau F (2003) The logic system of concept graphs with negation. Springer-Verlag, Berlin
9. Fong B, Spivak D (2019) Graphical regular logic. arXiv:1812.05765v2
10. Mac Lane S (2010) Categories for the working mathematician, 2nd edn. Springer, New York
11. O'Shea JR (2007) Wilfrid sellars: naturalism with a normative turn. Polity, Cambridge
12. Pietarinen A-V (2006) Signs of logic: Peircean themes on the philosophy of language, games and communication. Springer, Dordrecht
13. Roberts D (1973) The existential graphs of Charles S. Peirce. Mouton, The Hague
14. Selinger P (2010) A survey of graphical languages for monoidal categories. In: Coecke B (ed) New structures for physics. Springer, Heidelberg
15. Sellars W (1996) Naturalism and ontology. Ridgeview, Atascadero
16. Sellars W (1992) Science and metaphysics: variations on Kantian themes. Ridgeview, Atascadero
17. Shin S-J (2002) The iconic logic of Peirce's graphs. MIT, Cambridge
18. Wittgenstein L (2000) Tractatus logico-philosophicus. Routledge, London Trans. Ogden CK
19. Wittgenstein L (1978) Philosophical investigations. Basil Blackwell, Oxford Trans. Anscombe GEM
20. Zalamea F (2012) Peirce's logic of continuity. Docent, Boston



An Inferential View on Human Intuition and Expertise

Rico Hermkes(✉) and Hanna Mach

Goethe University, Theodor-W.-Adorno-Platz 4,
60629 Frankfurt/Main, Germany
hermkes@econ.uni-frankfurt.de,
mach@em.uni-frankfurt.de

Abstract. There are two central assumptions in expertise research. First, skilled performance and expertise are not genuinely deliberative processes, but closely related to intuition. Second, intuitions cannot be reduced to mere emotionally driven behaviour. Rather, they are considered as the acquisition and application of (tacit) knowledge. To distinguish it from deliberatively acquired knowledge, tacit knowledge is referred to as *know-how*. However, little is known about the logicality of these cognitive processes and how such *know-how* is acquired and applied *in actu*. The aims of this paper are (1) to explicate the cognitive characteristics of intuitive processes and (2) to propose a framework that enables us to model intuitive processes as inferences. For the first aim, we turn to Polanyi's theory of tacit knowing. For the second aim, we draw on Peirce's conception of abduction. In this respect, we shall consider the function of epistemic feelings for the validation of abductive results. Finally, we draw on the inferential approach developed by Minnameier, which integrates the Peircean inferences (abduction, deduction, induction) into a unified framework. This framework is used to explain how to proceed from abduced suggestions to tacit knowledge. As a result, we can explicate the inferential processes underlying intuition in detail. Expertise research might benefit from our findings in two ways. First, they suggest, how *know-how* might be generated in *actu*. Second, they may contribute to educational research in facilitating the acquisition of expertise *from rules to skillful know-how*.

Keywords: Intuition · Expertise · Tacit inference

1 Introduction

Human cognition is materially characterized by intuition and tacit processes. On the one hand, intuitions play an essential role in our everyday activities. On the other hand, they are particularly important for skilled performance and expertise. For example, we may think of the expertise of chess-players, who need to have an immediate understanding of the current game situation. We may envision economists, who make financial decisions under uncertainty. Or, we may think of teachers, who are engaged in the business of classroom management busy coordinating plenty of tasks simultaneously. In this respect, intuitions are of central interest in the research on expertise.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 274–286, 2019.
https://doi.org/10.1007/978-3-030-32722-4_16

Prominent examples in this field are the approach by Schön [1] called "reflective practitioner" and the stage model published by Dreyfus & Dreyfus [2], which describes professional development as a path "[f]rom rules to skillful know-how". According to Dreyfus & Dreyfus [2] individuals undergo five stages on their way to expertise, beginning with explicit rule-following at the stages of novices and of advanced beginners. In their model, the capacity of deliberative thinking and competent acting has reached its highest level at stage 3. At stage 4 (proficient) intuition and tacit processes attain increasing relevance. The final 5th stage is characterized by an intuitive performance which is expressed by the term know-how. Know-how already suggests that cognitive processes are involved in intuition. If this were true, intuitions should no longer be reduced to instinctive or emotionally driven behaviour, but rather regarded as processes that underlie some kind of cognitive control, although this control differs from deliberate and reflective cognition. Accordingly, research on human intuition focusses on the logicality and rationality of intuitions (see [3–6]). Nevertheless, the cognitive modelling of such processes remains an open question.

Therefore, the aim of this paper is to gain a deeper insight into the tacit nature of intuitions and to carve out how intuitions can be captured by a cognitive framework. Such a framework would not only explicate the rationality of intuition, but contribute in facilitating the acquisition of *know-how*. In this context, two questions arise that build on one another.

The first question is: What are the cognitive characteristics of intuitive processes? For answering this question, we will draw on the theory of Michael Polanyi who assumes acts of tacit knowing as inferential processes. Although Polanyi developed a philosophical theory, it had a strong impact on various other disciplines. In the 1990s, Neuweg [7, 8] introduced the "Tacit knowing view" in the field of education based on Polanyi's work. Economic theories also build on Polanyi's theory, e.g. by addressing questions of human rationality (see [4, 9]) or expertise in the context of entrepreneurship and management (see [1, 10]). However, Polanyi [11] does not fully explicate what kind of inferences he has in mind, when talking about "the logical structure of tacit inference" (p. 96).

This leads us to the second question: Is there an inferential framework that explicates the characteristics of tacit inferences? For answering this question, we shall relate to Peirce and his conception of abduction. A crucial issue in this matter is the validation and cognitive control of such inferential processes: As inferences are fallible, how can the validity of tacit inferences be judged without relying on deliberative thinking and conscious control? In this context, we shall discuss the role of epistemic feelings in abduction. Finally, we shall examine how tacit knowledge (*know-how*) can be attained form abductive inferred results. For this, we will turn to the inferential approach of Minnameier [12, 13], who integrated the three Peircean inferences (abduction, deduction, induction) into a unified cognitive framework. This framework equally accounts for low-level perceptual processes as for high level scientific reasoning, and explains how to progress from suggestions to knowledge.

This paper is structured as follows: Sect. 2 introduces Polanyi's theory of tacit knowing and explicates how a cognitive modelling of intuition can be accomplished. In Sect. 3, we will present the Peircean inferential approach on abduction. Here, we will show that Polanyi's approach of tacit inferences can be successfully integrated into the Peircean framework. Section 4 asks for the validation of inferential outcomes without

relying on deliberative thinking and emphasizes the role of epistemic feelings in this matter. Section 5 examines how tacit knowledge (*know-how*) can arise from abductive outcomes.

2 Polanyi's Theory of Tacit Knowing

In his theory, Polanyi focusses on the tacit powers of the mind. To him, the main purpose of the human mind is to establish coherence in nature. As Sanders [14] puts it, it is the successful discovery of a hidden coherence in nature what Polanyi subsumes under the concept of intuition. This accounts for all scales of cognitive processing. Just as perceptual processes serve to provide a coherent picture of the experienced world, the aim of high-level scientific reasoning is to attain a coherent picture of the worlds' hidden structures. The main difference between higher and lower level processing is, that in scientific reasoning the individual has access to a more sophisticated repertoire of cognitive tools, such as linguistic formats and symbolic sign systems. However, the processes on either level are fallible and therefore dependent on an underlying logicality. Polanyi [15] states that "the logic of perceptual integration [of sensory data into a coherent whole] may serve […] as a model for the logic of [scientific] discovery" (p. 2). According to Polanyi [16], this process of integration can be modelled as an inferential process, which he calls "tacit inference" (p. 313).

The term "perceptual integration" in this quote already indicates that two different scales are involved in the process: a lower scale where (sensory) data are represented, that are subsequently used for the process of integration, and a higher scale where the result of this integration is being located, i.e. the coherent whole (a *gestalt* in the case of perceptual integration). Hence, Polanyi models the structure of tacit knowing as the relation between terms on these two scales. The coherent entity, which Polanyi calls the *focal object*, is being represented in the distal term. The totality of all internal instances, states and data the person is making use of subsidiarily in the course of inferring the focal object, represents the proximal term.

The general principle underlying this process of integration can be expressed as follows: By relating the proximal term and the distal term to each other, the epistemic subject makes use of the data available to her in order to infer a new quality beyond the available data. In doing so, the subject assimilates the external world into her internal structures thereby attaining a higher understanding of the world. The data used subsidiarily may originate from various resources. They can be sensory data, activated representations from the memory or representations of bodily states.

We would like to give an example to illustrate the process of tacit knowing. The example refers to an act of writing which can be regarded as a case of skillful performance. The task is to hold a pen and to write on a surface that has a texture, which is so far unknown to the writer. In this case, the pressure of the pen has to be adapted to the texture of the table or to the underlying pad. However, the texture of the table surface is not directly accessible to the writing subject, because her hand and fingers do not have direct tactile contact with it. The contact is only mediated through the pen. Hence, the subject has to generate a hypothesis concerning the appropriate writing

pressure. According to Polanyi, the general solution principle is to make use of the data already available (proximal) in order to attain a focal object, which is beyond the subjects' experience at that moment (distal). The solution is to exploit the sensory data resulting from the contact of the fingers with the pen (proximal) in order to generate a hypothesis about the unknown texture of the pad, where the contact point between the tip of the pen and the pad is located (distal). The inferred hypothesis about the hardness of the surface is needed to choose an appropriate writing pressure. This example shows us that a simple task like holding a pen with an appropriate pressure involves an underlying cognitive process that is hypothesizing, essential creative and in fact far from trivial. As a result, the subject does not only acquire a practical skill. She also attains a higher and more abstract understanding of the world. In this case, she obtains a concept of *hardness* of an object, the surface, being an inherent part of the distal term.

Summing up Polanyi's conception, we may conclude that intuition – regarded as acts of tacit knowing – can be characterized as follows: (1) It should be regarded as an inferential process. (2) Such inferences are thoroughly creative: A new quality is achieved, that surpasses the subsidiaries in its meaning. (3) It enables the subject to comprehend the elements in their joint function, which would not be intelligible if standing by themselves. Polanyi [15] illustrates this issue using the example of a machine: "Viewed in themselves, the parts of a machine are meaningless; the machine is comprehended by attending from its parts to their joint function, which operates the machine" (p. 15). However, some important questions remain an open issue in Polanyi's conception. As Neuweg [7] correctly points out, Polanyi never fully explicates the logic of tacit inferences. To shed more light on Polanyi's "black box" concerning the characteristics of tacit inferences, we will now turn to Peirce and his conception of abduction.

3 Pragmatistic Conceptions of Abduction

3.1 Peirce's Inferential Approach

As stated at the beginning of this article, Polanyi strongly emphasizes the tacit power of the mind for establishing coherence in nature. A main purpose of the theory of tacit knowing is to explain how the human mind operates in attaining a coherent understanding of the world. The same issue can be found in the Peircean conception of abduction. It also includes the question of how to integrate several items into a new coherent entity and the related aspect of creativity in human thought. According to the function of abductive reasoning, Peirce [17] states: "Abduction […] is the only logical operation which introduces any new idea" (CP 5.171). He further asserts: "The abductive suggestion comes to us like a flash. […] It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together." (CP 5.181).

If we assume Polanyi's tacit inferences as abductive inferences, Peirce moves us a step further by explicating the logicality of abductive inferences. On the one hand, he develops a semi-formal scheme for abduction. In formalizing, abduction is depicted as a fallible process instead of a mere guess (see [18]). On the other hand, Peirce specifies

this inferential process by distinguishing three sub-processes or steps, which he calls *colligation*, *observation* and *judgment* (CP 2.442–2.444; see [13]).

Colligation consists of bringing together certain items that serve as premises "into one field of assertion" ([19], p. 45), needed for the upcoming step of observation. The observation step accounts for merging these items, which are finally converged into a conclusion. The final judgment step concerns the adoption or rejection of the inferred conclusion. The processing of subsidiaries into a focal object, that Polanyi describes, applies to the observational step, that is, the transition from premises (data) into a conclusion. The content of the proximal term consists of the data that are colligated for the subsequent [tacit] integration.

What is missing in Polanyi's conception is the judgment of validity of the inferred conclusion, that corresponds to the Peircean judgement step. Although Polanyi introduces 'coherence' as a criterion for valid inferences, he does not explicate how the mind succeeds in differentiating between good and bad inferred results *in actu*.

The abduction schema developed by Peirce [17] includes the missing judgment step and allows us to specify Polanyi's conception concerning this gap. The Peircean schema reads as follows:

(1) The surprising fact, C, is observed.
(2) But if A were true, C would be a matter of course.
(3) Hence, there is reason to suspect that A is true. (CP 5. 189)

Surprise is conceptualized as something that signals the need for explanation. Peirce [19] states:

> "The whole operation of reasoning begins with abduction. […] Its occasion is a surprise. That is, some belief, active or passive, formulated or unformulated, has just been broken up. […] The mind seeks to bring the facts, as modified by the new discovery, into order; that is, to form a general conception embracing them." (p. 287).

In the context of Polanyi's theory, the occurrence of the surprising fact C can be interpreted as the epistemic task that the mind has to solve when interacting with the world. This task consists in generating a coherent unity from the single data at hand. The inferred focal object could then be interpreted as an explanation (that is the A in the scheme) for the data C. In the case of visual perception, it means that the visual *gestalt* explains the sensory data at hand.

The purpose of a coherent result is twofold. Facing the initial problem, it accounts for the elimination of the disturbance caused by the surprising fact. Facing future interactions with the world, coherence accounts for gaining a "more predictive control over one's world" ([20], p. 294). The two criteria set out by El Khachab [21], also apply in this context. In his essay "The logical goodness of abduction in C.S. Peirce's thought", he writes:

> "[A] good hypothesis, as obtained through abduction, has two main characteristics: first, it must *explain* the given facts; second, it must be susceptible to future *experimental verification*." (p. 163).

The first criterion, which is *explaining the facts*, addresses the initial problem. The second criterion of *susceptibility to empirical verification* is directed towards future

interactions with the world, pointing to the forthcoming inductive inference. As Peirce [17] puts it, induction "consists in […] observing those [empirical] phenomena in order to see how nearly they agree with the theory" (CP 5.170). In the case of deliberative thinking, the application of the two criteria is uncritical for the reasoner. The question remains, however, whether the subject can also implicitly apply those criteria.

Likewise, Peirce was ambiguous about the assertion, whether non-deliberative, unconscious processes can be thoroughly conceived as abductive inferences. Moreover, he was unsure at which cognitive level inferential processes were supposed to set in. On the one side, Peirce writes that abductions should be regarded as "controlled thinking". Consequently, he was very critical about the inferential nature of perceptual processes. Peirce [17] states: "But the content of the perceptual judgment cannot be sensibly controlled now, nor is there any rational hope that it ever can be" (CP 5.212). On the other side, he states in the first cotary proposition (CP 5.181), that everything in the mind has its origin in the senses. Given that sensory processes are indeed based on uncontrolled processes, it follows that all premises of deliberate thinking are doomed to be insecure, and that, as a matter of fact, all human knowledge would be grounded on sand (for a detailed argumentation see [22]). Peirce [19] tries to cope with this issue in dismissing perceptions as an inferential boundary case, an "extreme case of abductive inferences" (p. 227).

3.2 Magnani's Eco-Cognitive Approach

Approaches in the tradition of Peircean Pragmatism have widened the scope of abduction. In drawing on Peirce's works on semiotics, Magnani [23, 24] developed the "eco-cognitive model of abduction". Hence, inferences were no longer restricted to symbolic sign systems (sentential abduction), and could be realized by using iconic sign formats (model-based abduction). According to Magnani [23], model-based reasoning occurs "when hypotheses are instantly derived from a stored series of previous experiences" (p. 268). Moreover, he assumes a type of "non-theoretical" inferences which he names "manipulative abductions" that "happens when we are thinking through doing" (p. 39; see [25]). In focusing on *practical reasoning*, Park [26] concludes:

> "[A]s manipulative abduction can be construed as a form of practical reasoning, the possibility of expanding the scope of abduction is wide open" (p. 212).

If abductions are conceived in such broader sense, then tacit processes may be subsumed under an inferential theory. However, assuming inferences at lower cognitive levels entails further challenges. These concern the two validity criteria for abductive inferences explicated by El Khachab [21]. The first challenge relates to the judgment of an inferred result: How can an epistemic subject prove the validity of a (tacit) inference without performing a deliberative judgment? The second challenge relates to the dignity of tacit knowledge. It refers to the criterion of *susceptibility to empirical verification*. If intuitive processes were restricted to abduction, no knowledge could be acquired at all. As Peirce clearly points out, abductions only lead to suggestions, not to knowledge. However, as stated above, research on expertise and human

intuition assumes that intuition is characterized by *know-how*. Thus, we need to explain how the subject may accomplish (tacit) knowledge from mere suggestions.

Polanyi seemed to be aware of that problem, too. In [27] he calls for the confirmation of the inferred focal object. He writes: "a formal step can be valid only by virtue of our tacit confirmation of it." (p. 131). Such a confirmation addresses the inductive testing of a previously abduced hypothesis or explanation.

In the next two sections we will attempt to meet those two challenges. In Sect. 4 we shall introduce the approach of Proust [28] which deals with the role of epistemic feelings in cognitive processes, particularly focusing on their suitability for judging the validity of cognitive outcomes. In Sect. 5 we will present the inferential framework developed by Minnameier [12], which describes how a reasoner can proceed from abduced suggestions to (tacit) knowledge. Coming back to our initial two questions, we will finally discuss our results in the light of expertise research.

4 Epistemic Feelings and Their Function in Tacit Inferences

The first question we need to answer is, how abductive inferences can be validated without engaging in deliberative thinking. Peirce and Polanyi both emphasize the significance of feelings in this matter. In pursuing Peirce's assumption, that surprise motivates/initiates the search for a new explanation, we are inclined to speculate that feelings might be involved in the validation of the abductive result. Polanyi [29] also indicates such a solution, stating that, tacit integration leading to a (new) coherent entity is accompanied by "feelings of a deepening coherence" (p. 7).

Research on human intuition tells us, that intuition and feelings are strongly entwined (see [6, 30–34]. Although feelings do not seem to correspond with the idea of inference at first sight, we shall argue that feelings are qualified candidates for monitoring inferential processes, concerning its course and its outcomes. Reflecting on creativity in cognition, Koriat [35] points out, "that the person can somehow subjectively monitor these underground processes" (p. 153) and that epistemic feelings may be involved in such processes. Monitoring relates to two aspects. First, it relates to the course of the inferential process. In this context, epistemic feelings serve to inform the agent whether a cognitive process is running fluently or whether it is stagnating (feeling of fluency). Second, monitoring relates to the judgement of its outcome, generally accompanied by feelings of rightness, coherence, certainty, etc. (see also [36], p. 701).

In referring to Koriat, Proust [28] developed a conception, where epistemic feelings fulfill either functions. Concerning the monitoring target, she discriminates between *predictive feelings* that address the course of the cognitive process and *retrospective feelings* that are related to outcomes. Concerning the latter, she states, that "their valence and intensity tell the agent whether she should accept or reject a cognitive outcome" (p. 6). Likewise, in the case of abduction, a feeling of coherence (or in the case of a negative valence of incoherence) informs the agent about the goodness or truth of an abductive outcome. In Peircean words, we may say, that the surprising fact that initiated the inferential process has been successfully explained away, now being a *matter of course*.

We may illustrate the validating function of epistemic feelings by presenting the example of an expert chess player who relies mainly on his intuition in the course of the game. Polanyi [16] describes such a situation as follows:

> "A chess player conducting a game sees the way the chess-men jointly bear on his chances of winning the game. This is the joint meaning of the chess-men to the player, as he decides from their position the choice of his next move." (p. 302f).

Let us now imagine that the opponent makes a move that the player immediately realizes as a severe attack. Technically speaking, a game constellation suddenly occurs that disturbs the coherent order of his play. In the context of chess, we may interpret coherence in terms of a mutual protection of one's own chess pieces and retaining the upper hand in the game.

Now it is the player's task to establish a new line-up of his pieces. She may spontaneously respond to the opponent's challenge by taking the piece that causes the disturbance. After doing so, the question arises whether the resulting game situation would be acceptable or not. There is no chance for the player to judge the outcome by checking all combinations of possible moves and countermoves as this would exceed the limited playing time and her mental capacities. This is the point where feelings come into play. Likewise, a feeling of coherence informs our player that the intended move might be suitable to overcome the disturbance and to re-establish a new coherent constellation on the chessboard. On the other hand, a feeling of uncertainty could signal the player that the intended move might not be sustainable enough for regaining superiority over a longer time-period. Only in the very next turn, the attack may be repeated, leaving our player so much the worse. If she nevertheless sticks to the intended move, we may say that the player acts against her intuition. Although the feeling of uncertainty strongly suggested her to reconsider the move, she might stick to her decision, e.g. due to a lack of alternatives or of time pressure.

The situation clearly illustrates the significance of epistemic feelings in the judgment of cognitive outcomes. First, they signal whether a result should be accepted or rejected, i.e. informing the individual about the validity of the result. Second, feelings either prompt the individual to carry on or to reconsider an outcome that might have been accepted too hasty or even against one's intuition.

However, the acceptance or rejection of the abduced result does not mark the endpoint of the whole process. Even if the result was judged as coherent, it needs to prove its worth in the empirical world. According to the criterion formulated by El Khachab (see Sect. 3.1), a step of empirical validation is mandatory. Only then would an individual be able to speak of the acquisition of *know-how*.

5 Modelling Tacit Processes as an Inferential Triad

Based on Peircean Pragmatism, Minnameier [12, 13] developed a framework that tells us how a subject can proceed from an abducted hypothesis to knowledge. According to his framework, this epistemic process – initiated by the surprising facts and finally leading to acquired knowledge – can be modelled as an inferential triad: (1) abduction leads to an explanatory hypothesis, (2) deduction draws predictions from this hypothesis

which are to be (3) tested inductively. And only if a positive inductive result is attained, we may speak of knowledge. In adapting the inferential triad to tacit inferences, the cycle can be described as follows: The first step accounts for inferring a focal object. The focal object may be, for example, a perceptual *gestalt* or a situational model. This step refers to abduction, which has been explained in detail earlier in this text, in referring to Polanyi's theory (Sect. 2) and the Peircean approach (Sect. 3).

The second step is about deriving implications from the focal object. These may be, for example, predictions about the further course that the situation will take: about possible outcomes or expected events. This step refers to deduction. Neuroscientific approaches called "predictive coding" emphasize this idea of anticipating future events and environmental states (see [37]). A major claim of these approaches is that the brain is not passively waiting for stimuli to impinge on it, but that it is actively making inferences all the time. That is why brains are coined "ever-active prediction engines" ([38], p. 2) in this framework. Moreover, some approaches suggest that predictive processing also accounts for the functioning of the mind (see [39–41]. Hawkins [39] gives us some vivid examples in his "Memory Prediction Framework" (p. 6). The examples concern skilled performance as well as everyday actions. Hawkins [39] states:

> "Every time you put your foot down while you are walking, your brain predicts when your foot will stop moving and how much "give" the material you step on will have. […Or] When you listen to a familiar melody, you hear the next note in your head before it occurs" (p. 62).

Predicting what comes next also explains why we quickly realize when a musician is out of tune, even when the piece is unknown to us. The reason for this is, that the focal object cannot merely be regarded as a "description of the melody". Moreover, it includes a general principle. What we expect is the unfolding of that principle as a melody. In unfolding, the melody generates empirical (sensory) data, which may subsequently get used as evidence for the abduced principle and its derived predictions.

The final third (inductive) step concerns the confirmation of the expected events derived in step two. Only if the focal object is confirmed by induction, it may be considered as tacit knowledge – or *know-how*. Whenever the focal object is falsified – which means that a prediction error occurred – the subject is prompted to infer a new focal object. Consequently, the inferential cycle starts again. In this respect we are reminded that prediction errors should not only be regarded as something 'wrong', but also as an impulse for learning processes and the attainment of a higher skill level.

Minnameier's inferential approach allows us to classify inferential processes leading to different forms of *know-how*. Minnameier [42] does not constrain the inferential cycle to the explanatory domain, but extends it to the technological and ethical domains. Depending on the domain, a specific validity criterion applies to the final inductive inference that leads to knowledge: (1) *truth* for the explanatory domain, (2) *effectivity* for the technological domain, and (3) *justice* for the ethical domain. Perceptual and diagnostic processes, aiming at the integration of sensory data, are subsumed under the explanatory domain. Skilled performance, like playing chess or writing on an unknown surface, however, are located in the technological domain. Practical reasoning in this domain is not directed at truth but on effectiveness in achieving a goal. The expertise of teachers in classroom management or of a carpenter

in building an attic both belong to this domain, too. Moral reasoning and moral intuition are assigned to the ethical domain, its validation criterion being justice (e.g. fairness; see [43]).

Moreover, an extension of the classification for another domain seems conceivable. When thinking about the *know-how* of musicians, composers, painters, poets, dancers, etc., we are inclined to add *aesthetics* as a forth domain. A manifold of scientific work focusses on skillful performance and *know-how* in aesthetics (see [1, 44]. An appropriate validation criterion for the inferences related to this domain might be *beauty*.

Putting these things together, we may obtain an inferential framework that is suitable for modelling *tacit inferences*. This is not restricted to the explanatory domain but also accounts for skilled action in real-world environments aimed at effectiveness, justice or beauty. As the triadic approach also incorporates the final inductive testing, an empirical verification/falsification is feasible. Therefore, we can legitimately speak of the acquisition of tacit knowledge.

6 Conclusion

Expertise research emphasizes the importance of intuition and tacit processes. It is assumed that such processes cannot be reduced to mere emotionally driven behaviour, but should rather be considered as cognitive. The aims of this paper were (1) to explicate the cognitive characteristics of intuitive processes and (2) to propose a framework that enables us to model intuitive processes as inferences. In addressing the first aim, we drew on Polanyi's theory of tacit knowing. For meeting the second aim, we turned to Peirce and his conception of abduction. Dealing with the latter, we identified two challenges concerning the validity criteria for abductive inferences.

The first challenge pointed to the validity of an abduced result and how it could be proven without performing a deliberative judgment. We suggested epistemic feelings as possible candidates for judging whether an inferred outcome should be accepted or rejected. Epistemic feelings also invite a reasoner to carry on or to reconsider an outcome that might have been accepted too hasty or against one's intuition.

The second challenge dealt with the dignity of tacit knowledge. Even if an abductive outcome is judged as valid, its worth needs to be proven inductively. For this, we drew to Minnameier's inferential framework that models knowledge acquisition as an inferential cycle comprising abduction, deduction and induction. By means of this framework we may explain how the subject may proceed from abduced results to tacit knowledge (*know-how*).

In sum, we are able to model intuitions as processes leading to the acquisition of tacit knowledge. Moreover, we can explicate the inferential processes underlying intuition in detail. Our conception is in line with recent approaches which aim to downsize Kahneman's [3] separation into two functionally distinct cognitive systems (system 1, system 2) (see [45] for a critical review and [46] for experimental findings). For instance, Smith [4] does no longer separate rationality from intuition, but rather subsumes intuition under the umbrella of rationality. In referring to Polanyi, he distinguishes two forms of rationality, i.e. constructivist and ecological rationality. These two forms are "not inherently in opposition" (p. 2), but to be considered as interacting

with each other. Yet, some approaches go beyond this idea in arguing for an even more integrated cognitive architecture (see [38, 47]). On the horizon, this could result in a conception assuming a unified form of rationality.

Expertise research can benefit from our findings in two ways. First, they suggest, how *know-how* might be generated *in actu*. Second, they may contribute to educational research in facilitating the acquisition of expertise *from rules to skillful know-how*.

Acknowledgements. We would like to thank Gerhard Minnameier and Tim Bonowski for their helpful discussions.

References

1. Schön DA (1983) The reflective practitioner. how professionals think in action. Basic books
2. Dreyfus HL, Dreyfus SE (1986) Mind over machine: the power of human intuition and expertise in the era of computer. Free Press, New York
3. Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York
4. Smith VL (2008) Rationality in economics. Constructivist and ecological forms. CUP, New York
5. De Neys W (2012) Bias and conflict: a case for logical intuitions. Perspect. Psychol. Sci. 7:28–38. https://doi.org/10.1177/1745691611429354
6. Thompson V (2014) What intuitions are… and are not. In: Ross BH (ed) The psychology of learning and motivation, vol 60. Elsevier, San Diego, pp 35–75
7. Neuweg GH (2004) Könnerschaft und implizites Wissen, 3rd edn. Waxmann, Münster
8. Neuweg GH (2015) Tacit knowing and implicit learning. In: Neuweg GH (ed) Das Schweigen der Könner. Waxmann, Münster, pp 81–96
9. Hayek F (1978) New studies in philosophy, politics, economics and the history of ideas. Routledge & Keagan Paul, London
10. Raymond CM, Fazey I, Reed MS, Stringer LC, Robinson GM, Evely AC (2010) Integrating local and scientific knowledge for environmental management. J Environ Manage 91:1766–1777. https://doi.org/10.1016/j.jenvman.2010.03.023
11. Polanyi M (1968) The body-mind relation. In: Coulson WR, Rogers CR (eds) Man and the Sciences of Man. Merrill, Columbus, pp 85–102
12. Minnameier G (2005) Wissen und inferentielles Denken: Zur Analyse und Gestaltung von Lehr-Lern-Prozessen. Lang, Frankfurt am Main
13. Minnameier G (2017) Forms of abduction and an inferential taxonomy. In: Magnani L, Bertolotti T (eds) Springer handbook of model-based science. Springer, Berlin, pp 175–195
14. Sanders AF (1988) Michael Polanyi's post-critical epistemology. A reconstruction of some aspects of "tacit knowing". Rodopi, Amsterdam (1988)
15. Polanyi M (1966) The logic of tacit inference. Philosophy 41:1–18. https://doi.org/10.1017/S0031819100066110
16. Polanyi M (1967) Sense-giving and sense-reading. Philosophy 42:301–325. https://doi.org/10.1017/S0031819100001509
17. Peirce CS (1903/1960) Lectures on pragmatism. In: Hartshorne C, Weiss P (eds) Collected papers of Charles Sanders Peirce, vol 5 and 6. Belknap Press, Cambridge
18. Douglas, W.: Abductive reasoning. University of Alabama Press (2015)
19. Peirce Edition Project: The essential peirce. selected philosophical writings, vol 2. Bloomington, Indianapolis (1998)

20. Hawkins J, Pea RD (1987) Tools for bringing the culture of everyday and scientific thinking. J Res Sci Teach 24(4):291–307. https://doi.org/10.1002/tea.3660240404
21. El Khachab C (2013) The logical goodness of abduction in C.S. Peirce's thought. Trans Charles S. Peirce Soc 49:157–177. https://doi.org/10.2979/trancharpeirsoc.49.2.157
22. Hermkes R (2016) Perception, abduction, and tacit inference. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology, studies in applied philosophy, epistemology and rational ethics 27. Springer International Publishing, Cham, pp 399–418
23. Magnani L (2009) Abductive cognition. The epistemological and eco-cognitive dimension of hypothetical reasoning. Springer, Berlin
24. Magnani L (2016) The eco-cognitive model of abduction II. Irrelevance and implausibility exculpated. J Appl Logic 15:94–129. https://doi.org/10.1016/j.jal.2016.02.001
25. Magnani L (2004) Reasoning through doing. Epistemic mediators in scientific discovery. J Appl Logic 2:439–450. https://doi.org/10.1016/j.jal.2004.07.004
26. Park W (2017) Magnani's manipulative abduction. In: Magnani L, Bertolotti T (eds) Springer handbook of model-based science. Springer, Cham, pp 197–213
27. Polanyi M (1962) Personal knowledge. Toward a post-critical philosophy. corrected edition. Routledge, London
28. Proust J (2015) The representational structure of feelings. In: Metzinger T, Windt JM (eds) Open MIND: 31. MIND Group, Frankfurt am Main. https://doi.org/10.15502/9783958570047
29. Polanyi M The creative imagination. Wesleyan Lectures. Lecture 3. http://www.polanyisociety.org/WesleyanLectures/Weslyn-lec3-10-21-65.pdf. Accessed 2 Feb 2019
30. Westcott MR, Ranzoni JH (1963) Correlates of intuitive thinking. Psychol Rep 12:595–613. https://doi.org/10.2466/pr0.1963.12.2.595
31. Bastick CI (1982) Intuition: how we think and act. Wiley, New York
32. Schwartz N (1990) Feelings as information: informational and motivational functions of affective states. In: Higgins ET, Sorrentino RM (eds) Handbook of motivation and cognition: foundations of social behavior, vol 2. Guilford, New York, pp 527–651
33. Gigerenzer G (2008) Gut feelings: the intelligence of the unconscious. Penguin Books, New York
34. Epstein S (2010) Demystifying intuition: what it is, what it does, and how it does it. Psychol Inq 21:295–312. https://doi.org/10.1080/1047840X.2010.523875
35. Koriat A (2000) The feeling of knowing: some metatheoretical implications for consciousness and control. Conscious Cogn 9:149–171. https://doi.org/10.1006/ccog.2000.0433
36. McDermott R (2004) The feeling of rationality: the meaning of neuroscientific advances for political science. Perspect. Polit. 2:691–706. https://doi.org/10.1017/S1537592704040459
37. Friston K (2003) Learning and inference in the brain. Neural Netw 16:1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005
38. Clark A (2015) Embodied prediction. In: Metzinger T, Windt JM (eds) Open MIND: 7. MIND Group, Frankfurt am Main. https://doi.org/10.15502/9783958570115
39. Hawkins J (2006) On intelligence. Times Books, New York
40. Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–253. https://doi.org/10.1017/S0140525X12000477
41. Hohwy J (2013) The predictive mind. OUP, Oxford
42. Minnameier G (2016) Abduction, selection, and selective abduction. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology – logical, epistemological, and cognitive issues. Springer, Heidelberg, pp 309–318
43. Minnameier G (2016) Rationalität und Moralität. Zum systematischen Ort der Moral im Kontext von Präferenzen und Restriktionen. J Bus Econ Ethics 17(2):259–285. https://doi.org/10.1688/zfwu-2016-02-minnameier

44. Zembylas T, Niederauer M (2017) Composing processes and Artistic Agency. Tacit knowledge in composing. Routledge, London
45. Evans JS, Stanovich KE (2013) Dual-process theories of higher cognition: advancing the debate. Perspect Psychol Sci 8:223–241. https://doi.org/10.1177/1745691612460685
46. Garrison KE, Handley IM (2017) Not merely experiential: unconscious thought can be rational. Front Psychol 8:199–211. https://doi.org/10.3389/fpsyg.2017.01096
47. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RL (2011) Model-based influences on human's choices and striatal prediction errors. Neuron 69:1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027



Disseminated Causation: A Model-Theoretical Approach to Sophisticated Abduction

Andrés Rivadulla(✉)

Department of Logic and Theoretical Philosophy,
Complutense University of Madrid, 28040 Madrid, Spain
arivadul@ucm.es

Abstract. How does theoretical science implement the search for the best explanation of complex phenomena? Is it possible for these explanations to be causal? These are the two main questions that I intend to analyze in this paper.

In the absence of theories capable of offering an explanation of novel or surprising phenomena, science resorts to abduction in order to find the hypothesis that best accounts for the observations. Now, abduction is not the only way that makes explanation possible or supports scientific creativity. Theoretical physicists usually combine mathematically, in a form compatible with dimensional analysis, already accepted results proceeding from different branches of physics, in order to *anticipate/explain* new ideas. I propose the name of *theoretical prediction*, for this kind of reasoning.

Usually the *theoretical models* designed by physicists in order to offer an explanation of the observations are built by applying predictive reasoning. The explanation they provide is inter-theoretical. In these cases prediction comes in support of abduction, and since it is not *standard abduction* which is taking place here, I name this procedure *sophisticated abduction*. Thus, if the desired explanation should be causal, this procedure would require going back to other causes or mixing causes with each other. Causation would be *disseminated* in a network of nomological chains.

Keywords: Causal explanation · Theoretical explanation · Theoretical models · Theoretical prediction · Sophisticated abduction · Disseminated causation

1 Introduction

To have a good explanation of what we see, even more of what surprises us, seems to be the greatest aspiration of every philosopher, every scientist, and anyone interested in the world in which we live. This is the cultural paradigm in the West. But, how does theoretical science implement the search for the best explanation of complex phenomena? Is it possible for these explanations to be causal? These are the two main questions that I intend to analyze in this contribution.

To paraphrase Rorty we can say that since the time of Aristotle we philosophers are heirs of a tradition according to which one of the fundamental goals of science is to *explain the phenomena*. Aristotle himself claimed in the *Posterior Analytics* that we believe that we know something when we know the cause. Since then we have

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 287–297, 2019.
https://doi.org/10.1007/978-3-030-32722-4_17

accumulated a considerable amount of theoretical knowledge, and this often allows us to find the desired explanation within the framework of the *best available theory*. We speak in this case of *theoretical explanation* and Popper-Hempel's *D-N* model offers us the guideline on the matter. Thus, the Newtonian celestial mechanics provides a theoretical explanation of Kepler's empirical laws; Planck's radiation law offers a theoretical explanation of both Stefan's black body radiation law and Wien's displacement law; Bohr's atomic model gives the first theoretical explanation of the elements' spectra, etc.

But the question that inevitably arises is whether there is *always* a theory available, or if sometimes this does not happen. When the latter occurs, instead of ignoring the facts, scientists implement some forms of ampliative reasoning, beginning with induction, which goes back to Aristotle himself. And, when it comes to surprising or novel facts, Charles Peirce unveiled, at the beginning of the twentieth century, the logical scheme of a form of reasoning, ampliative as well, to which he gave the name of *abduction*. When, in the middle of the 20th century, Gilbert Harman (1965) renamed abduction *inference to the best explanation*, the philosophers of science recognized that many of the great advances of science in the past had been the result of the application of abductive reasoning, rather than induction. Lipton (2001: 56) for instance assumes that "scientists infer from the available evidence to the hypothesis which would, if correct, best explain that evidence". And Josephson, Magnani and others recognize that abduction has two sides: inferential and explanatory; abduction generates plausible hypotheses and provides best explanations of facts, Magnani (2001: 17–18, 2007: 294) claims.

Indeed, we only have to look inside the history of natural sciences to realize that abduction has been widely applied since the beginning of Western science. I have presented elsewhere examples of abductive reasoning both in *observational sciences*: The postulation of *Homo antecessor* as a new hominin species, and the continental drift hypothesis; and in *theoretical sciences*: The dark matter and dark energy hypotheses, among many others.

Now, abduction is not the only way that makes explanation possible or supports scientific creativity. Theoretical physicists usually combine mathematically, in a form which is compatible with dimensional analysis, already accepted results from different branches of physics, in order to *anticipate* new ideas. The results postulated *methodologically* as premises proceed from differing theories, and any accepted result can serve as a premise – on the understanding that *accepted* does not mean *accepted as true*. Thus I maintain that in the methodology of theoretical sciences, like physics, we can implement deductive reasoning for theoretical *innovation* or *creativity* and I propose the name of *theoretical prediction*, or simply *prediction*, for these purposes. It is true, that Peirce (CP: 5.145) claims: "Induction can never originate any idea whatever. No more can deduction" and that "deduction merely evolves the necessary consequences of a pure hypothesis" (CP: 5.171). But in this Peirce seems not to be entirely correct.

The fact that the premises of the predictive argument proceed from different theories makes from prediction a *transverse* or *inter-theoretical* form of deductive reasoning. This is what makes it possible to *anticipate* new ideas in physics, i.e. to support *theoretical innovation*. Prediction differs fundamentally from abduction by the fact that

the results of preductive reasoning do not proceed from empirical data but they are deductively derived from the available theoretical background taken as a whole.

In theoretical physics very often hard physical-mathematical work is needed when the empirical data do not readily suggest a 'spontaneous' innovative explanation, as it was the case of Rutherford's planetary atomic model or Wegener's continental drift hypothesis. In those situations preductive reasoning comes in support of abduction: the theoretical explanation takes place *more preductivo*, i.e. the inference to the best explanation depends on the implementation of preductive reasoning, and since it is not *standard abduction* which is taking place here, I name this procedure *sophisticated abduction*. In Rivadulla (2016), for example, I claim the existence of this form of reasoning in theoretical physics, and I have even listed some phenomena whose explanation instantiate *sophisticated abduction*.

One question that can arise now is what relationships exist, if any, between sophisticated abduction and causal explanation. In Rivadulla (2019) I argue that the existence of inter-theoretical incompatibility makes it impossible at times to give a causal explanation of relatively simple facts, such as the fall of bodies or the planetary movements. The more for those phenomena that require for their explanation to resort to sophisticated forms of abduction!

Faced with the impossibility of offering imaginative, spontaneous, 'causal' hypotheses, celestial bodies and phenomena like stellar interiors, stellar atmospheres, novae, supernovae, white dwarfs, pulsating stars, neutron stars, pulsars, black holes, etc. whose internal processes are not observable, would be left unexplained. Instead, to deal with them astrophysicists design *theoretical models*. These models assume the explanatory role of abductive hypotheses and they are intended to reproduce the observed behaviour of the objects investigated.

These models proceed from the combination of accepted results of most of the theories and disciplines of physics, i.e. by applying *preductive reasoning*. The explanation they provide is therefore inter-theoretical. Thus, if the desired explanation should be causal, this procedure would require going back to other causes or mixing causes with each other, that is, it should reveal the existence of *nomic causal chains*. Since the explanation of an event might imply the concurrence of several explanatory laws, that is, a concatenation of explanations, then *causation would be disseminated*. The idea of a clear and distinct causal explanation would be completely blurred.

Theoretical explanation by means of theoretical models can not be expected to be one hundred percent correct, i.e. to be a *causal* explanation properly. But this is all we can expect. All that is in our hands is to provide explanations through the *preductive* construction of theoretical models, which supports *sophisticated abduction*.

In conclusion: Scientific explanation of complex phenomena by theoretical models can be taken as an implementation of sophisticated abduction. But the question about the cause(s) of the phenomena only admits in response that *causation is disseminated* in nomic chains. Moreover, we can not pretend that a theoretical model gives us the truth, the whole truth and nothing but the truth about the phenomena investigated.

In Sect. 2 I make a brief historical review until the seventeenth century, intending to identify the historical moments in which scientists are personally involved in the search for explanatory causes of physical phenomena, and then I connect with the conviction, by many philosophers from Newton onwards, that science can not be

restricted to the search of *proximate causes*, suggesting that *causality may be distributed or disseminated in causal chains*.

In Sect. 3 I focus on the white dwarf model in order to instantiate that a theoretical explanation of this kind of celestial objects can only be developed in a preductive way, i.e. resorting to the combination of a network of previously available theoretical results of various theories. As the conclusion will be a theoretical explanation of the observations that this kind of objects provide, I conclude that this theoretical model provides a kind of *sophisticated* abductive explanation. Finally, if we pretend that this explanation should in some way be causal, we must conclude that causation or causality, such as the explanation of a white dwarf, is distributed or disseminated in the network that the preductive construction of the theoretical white dwarf model makes possible.

2 Proximate Causes and Causal Chains

The scientist who, for the first time in the history of the West, offers a causal explanation of a physical phenomenon is Aristotle, in the fourth century before our era. As a good observer, Aristotle knows the existence of the moon eclipses and wonders about their cause.

In the *Posterior Analytics* 90ª15 Aristotle (1975) claims: "What is an eclipse? Privation of light from the moon by the earth's screening. Why is there an eclipse? or Why is the moon eclipsed? Because the light leaves it when the earth screens it." Thus Aristotle finds out the *immediate* or *proximate cause* of the moon eclipses.

In the thirteenth century the West recovers much of the Aristotelian works and since then the main task of the scholastic philosophers was to comment on the work of the Stagirite. But when the Copernican revolution begins, the question about the cause of planetary movements becomes peremptory.

This question is already present in Copernicus's *Commentariolus*. In the Praefatio, Copernicus affirms: "I observe that our ancestors assumed a large number of celestial spheres as a cause in order to regularly save the apparent motion of the planets." A glimpse of the idea of causation is also given by Copernicus in the Assumptions of his *Commentariolus*; Sexta petitio: "What appear to us as motions of the sun arise not from its motion but from the motion of the earth and our sphere, with which we revolve about the sun like any other planet." (Rosen 1959: 58–59), and Septima petitio: "The apparent retrograde and direct motion of the planets arises not from their motion but from the earth's. The motion of the earth alone, therefore, suffices to explain so many apparent inequalities in the heavens." (Rosen 1959: 59)

In the Preface to Copernicus's *De Revolutionibus* Osiander affirms: "it is the duty of an astronomer to compose the history of the celestial motions through careful and skilful observation. Then turning to the causes of these motions or hypotheses about them, he must conceive and devise, since he cannot in any way attain to *the true causes*, such hypotheses as, being assumed, enable the motions to be calculated correctly from the principles of geometry, for the future as well as for the past." (Rosen 1959: 24)

In chapter XXII of his *Mysterium Cosmographicum*, entitled "*Why a planet moves uniformly around the centre of the equant*", Johannes Kepler advances some ideas of a

dynamic approach to planetary movements. He presents the *cause* of delay and rapidity in the orbs of each of the planets (causa tarditatis et velocitatis in singulorum orbibus) in the following way:

> the planet will be slower because it departs more from the Sun and is moved by a weaker force (quia longius a Sole recedit et a virtute debiliori mouetur); in the rest it will be faster because it is closer to the Sun and under greater force (quia Soli vicinior et in fortiori virtute).[1]

And in Note 1 of this chapter Kepler states that the cause that Saturn, which is farther from the Sun, is slower than Jupiter, which is closer to the Sun, is the same that causes Saturn, in its aphelion, to be slower than in its perihelion: "The cause of both things is the greater or lesser straight elongation of the planet with respect to the Sun, because when it is farther from the Sun it moves with a slighter and weaker solar force." (Causa utriusque rei est elongatio Planetae a Sole rectilinea maior vel minor, quia longe distans a Sole versatur in virtute Solari tenuiore et imbecilliore.")

In 1609 Kepler publishes *ASTRONOMIA NOVA* ΑΙΤΙΟΛΟΓΗΤΟΣ*, seu PHYSICA COELESTIS* (*New astronomy based upon causes, or celestial physics*). According to Alexandre Koyré (1961: 159), Kepler dedicates his *Astronomia Nova* "to specify the nature of the force that moves the planets, more precisely to replace the 'animic' force of the *Mysterium Cosmographicum* by a physical force, quasi-magnetic, and to determine the strict mathematical laws that govern its action, as well as to elaborate a new theory of planetary motion, on the basis of the observational data provided by Tycho Brahe's work."

Indeed, in the Introduction to his *Astronomia Nova* Kepler (1992: 48) affirms: "I inquire into celestial physics and the natural causes of the motions." And he insists:

> the source of the five planets' motion is in the sun itself. It is therefore very likely that the source of the earth's motion is in the same place as the source of the other five planets' motion, namely, in the sun as well. It is therefore likely that the earth is moved, since a likely cause of its motion is apparent. (Kepler 1992: 52)

On the cause of this movement Kepler (*op. cit.*: 350–351) claims:

> What if the bodies of the planets were round magnets? As regards the Earth (one more planet, according to Copernicus) there is no doubt about it. William Gilbert[2] proved it. But this property must be described more precisely, for example by saying that the planet's globe has two poles, one of which follows the Sun while the other runs away from the Sun.

The magnetism would explain the movement of the stars, and in particular the rotation of the Earth. The source – the cause – of the motion of the planets would be in the Sun itself.

Galilei's *Dialogues Concerning Two New Sciences*, 1638, is a work dedicated to the investigation of the causes of physical phenomena. For instance, the vacuum is the cause of the adhesion between two plates, the resistance of the medium explains the variations of speed observed in bodies of different weights "so that if this is removed all bodies would fall with the same velocity" (Galilei 1954: 73), the vibrations of a string

[1] The Latin quotations come from Hammer (1963: 121).

[2] Gilbert (1544–1603) had published in London in 1600 *De Magnete*, where he states that the Earth is a magnet.

or the friction of a glass cup are the cause of the vibrations of the air and therefore of the sound (Galilei 1954: 98), the gravity is the cause of the acceleration of the heavy bodies (Galilei 1954: 165), etc. The philosophical premises of Galilei's investigations are: (1) "The cause must precede the effect"; (2) "every positive effect must have a positive cause"; and, following Aristotle, (3) "the non-existent cause can produce no effect" (Galilei 1954: 12).

Robert Boyle (1627–1691) and Robert Hooke (1635–1702) contributed considerably to the progress of experimental physics during the seventeenth century. The first one is known above all by the law that bears his name: At constant temperature the volume occupied by a gas is inversely proportional to the pressure to which it is subjected. And the second is the author of the law, which also bears his name, according to which the deformation experienced by an elastic material is proportional to the deforming force. In 1686 Edmond Halley (1656–1742) published an "Historical Account of the Trade Winds, and Monsoons" with an attempt to assign the *Physical Cause* of the said winds, and in 1735 George Hadley (1685–1768) also published an article entitled "Concerning the *Cause* of the General Trade-Winds" (My emphasis, A.R.). A century later, William Ferrel (1817–1891) applied the Coriolis effect in Ferrel (1856) to propose his hypothesis on the cause of winds and ocean currents. And two centuries later, Berzelius (1779–1844) also investigated in 1812/1813 the causes of the chemical proportions of elements in chemical compounds. That is, from the seventeenth century science was entrusted with the task of looking for the causes of phenomena.

From Isaac Newton onwards, one of the main questions that the philosophy of science deals with is whether celestial mechanics offers a causal explanation of the planetary motions. This question is part of the discussion around both the very idea of causality and the knowability of causes, which from Bacon, Berkeley and Hume to our days, with alternating optimistic and sceptical positions over the centuries, occupied the epistemologists. But, as I have already dealt with this issue in Rivadulla (2019), I will not enter it here.

As science progresses, it is evident that it does not merely conform to the explanation of the *immediate* or *proximate* cause of the observations, but rather it looks for hidden, mediated, second-level causes and this makes it possible to foresee that causation may be diluted in *networks of causal chains*, so that the more theoretical a science becomes, that is, the greater the weight of the theoretical component(s) is, the less able it is to clearly establish where things come from.

Already Newton himself manifests in different letters to Richard Bentley (1662–1742) his scepticism about the knowability of the cause of gravity. This shows that gravity had to obey itself to the action of another agent. In his *Optics* Newton (1782: 263) distinguishes between *particular causes* and *general causes*, and all this leads to think in the existence of *causal chains*, although, of course, Newton does not raise the issue in these terms. But retaking Newton's concern about what is the cause of gravity, William Whewell (1847: 434) speaks explicitly of succession of causes: *proximate causes* and *further causes*. And David Lewis (1986: 214) claims that the explanation of an event can be found "At the end of a long and complicated *causal history*."

In short, if we want science to reveal something more than the *immediate* or *proximate causes* of things, then we are obliged to take into account what I call *causal dissemination in nomic chains*. With this I mean both the idea of a 'linear' succession

in causal chains and the existence of *causal networks*, the last ones more typical of the theoretical explanations of complex phenomena.

3 Sophisticated Abduction and Disseminated Causation. The Theoretical *White Dwarf* Model

A theoretical model is a successful tool if it incorporates *necessary physical conditions* that facilitate the explanation of the observations, i.e. if it gives account of, saves, recovers or reproduces the phenomena in a satisfactory way. It can not be expected, however, that they constitute *sufficient conditions* to faithfully reproduce the observations. On the other hand, the greater the number of necessary conditions that make up the theoretical model, the clearer it becomes that causality, and therefore the causal explanation, is *disseminated* among the conditions interacting in an interlaced way.

Is there an explanation possible that we could accept as the *best explanation* at all of the observational characteristics of white dwarfs? A theoretical model of white dwarf has to offer a *theoretical explanation* that makes the existence of this kind of celestial objects credible. What theoretical ingredients we must incorporate into this model is the subject of this section.

The brightest 'star' of the night sky, after the Moon, Jupiter and Venus, is Sirius. This fact is due to its closeness to Earth: 8.65 light years. Its name is *Alfa Canis Maioris* (*α CMa*) and it is easily identifiable by extending to the southeast the line of the three main stars of the Orion belt. With Procyon and Betelgeuse it makes up the so-called *winter triangle* of the northern hemisphere.

Being the brightest 'star', it inevitably awakened, since the dawn of mankind, the interest of observers and astronomers. Suffice it to say that, for the Egyptians, the first annual appearance of Sirius in the night sky marked the beginning of the flooding of the Nile.

A surprise occurred when in 1844 the German astronomer Friedrich Wilhelm Bessel (1784–1846) claimed "I find, namely, that existing observations entitle us without hesitation to affirm that the proper motions, of *Procyon* in declination, and of *Sirius* in right ascension, are not constant; but, on the contrary, that they have, since the year 1755, been very sensibly altered." (Bessel 1844: 136) Excluded instrumental and calculation errors, Bessel (1844: 139) stated: "I have investigated the conditions which must be fulfilled, that a sensible change of the proper motion, like that observed, may be capable of explanation by means of a force of gravitation." After considering four hypotheses, among them the existence of an attracting star S_n located at a distance r_n from the star S that shows the anomalous behavior, Bessel (1844: 141) concludes – by means of a way of reasoning that I call *abduction by elimination of alternative hypotheses*[3] – that "Stars, whose motions, since 1755, have shewn remarkable changes, must (if the change cannot be proved to be independent of gravitation) be parts of smaller systems. If we were to regard *Sirius* and *Procyon* as double stars, the change of

[3] I have tackled this form of abduction by elimination of alternative hypotheses in Rivadulla (2008: 129–130; 2015: 146–147 and 2018: 70–71).

their motions would not surprise us; we should acknowledge them as necessary, and have only to investigate their amount of observation." This argument by Bessel, which anticipates the *logical scheme* of Peirce's abductive reasoning (CP: 5.189), opts for the explanatory hypothesis "That r_n be small, that is, the attracting mass very near to the disturbed star". Thus, Sirius must be a binary star system. From a causal point of view it would be necessary to conclude that the *proximate* or *immediate cause* of the 'anomalous' behaviour of Sirius would be the gravitational attraction exerted by another star, hitherto unknown.

Bessel's reasoning, which bets on the hypothesis of the existence of a celestial body that gravitationally disturbs the orbit of Sirius, is acceptable and the further development of astronomy would prove it right. This way of reasoning, akin to what in methodology of science is known as *ad hoc* hypotheses, was applied by Le Verrier (1811–1877) in 1845 as he pointed to the existence of the planet Neptune; but it was not successful when Le Verrier himself postulated the existence of Vulcan, a planet whose hypothetical gravitational interaction with Mercury would explain the 'anomalous' perihelion of this planet.

But, as I say, in the case at hand, Bessel was right. The astronomer Alvan Graham Clark (1832–1897), without intentionally intending – a typical case of pseudo-serendipity –, discovered in 1862 the companion star to which he gave the name of *Sirius B*. Indeed, Sirius was not really a star but a binary system consisting of two stars, *Sirius A* and *Sirius B* with masses $M_A = 2.3\ M_\odot$ and $M_B = 1.053\ M_\odot$ respectively. Later, Walter Sidney Adams (1876–1956), director of the Mount Wilson Observatory, discovered that *Sirius B*, with a radius of only 5.5×10^8 cm, has a surface temperature of 27,000 K, much hotter than *Sirius A* which has a temperature of 9,910 K. Thus *Sirius B* is a white dwarf, a kind of stars with approximately the mass of the Sun and the size of the Earth! Astrophysicists faced the task of explaining how objects of this type can exist.

Only a *preductive* process of reasoning can facilitate the construction of a *white dwarf theoretical model* compatible with observations, a model that claims the maximum value of the mass that can be supported by the internal pressure of a white dwarf. Well then, the preductive process in question takes place combining results from:

- Newtonian Mechanics: Hydrostatic Equilibrium Equation.
- Quantum Mechanics: Pauli's exclusion principle, Fermi energy for a completely degenerate electron gas (fermions), and Heisenberg's uncertainty principle.
- Theory of Relativity: limit speed c for degenerate electrons.

The construction *more preductivo* of a theoretical white dwarf model provides a *sophisticated abduction* of the explanatory hypothesis of the white dwarfs' observational data[4], a form of reasoning that will end up 'discovering' at least part of the *internal structure* of these stars.

To carry out this task we must face the problem of the pressure inside a white dwarf, a pressure capable of supporting its own weight in order not to collapse. It is, therefore, about investigating the physical conditions that make possible the balance between gravity and inner pressure.

[4] On the observed properties of white dwarfs see Hansen et al. (2004: 467–469).

Suppose that Sirius *B* is spherically symmetric and static. The hydrostatic equilibrium condition of *classical mechanic*, which gives the change in pressure with distance, allows us to calculate the pressure that Sirius *B* must support at its centre: a value corresponding to a million and a half times the pressure of the Sun in its own centre.

According to Chandrasehkar (1939: 412): "The white stars differ from those we have considered so far in two fundamental respects. First, they are what might be called 'highly underluminous'; that is, judged with reference to an 'average' star of the same mass, the white dwarf is much fainter. … Second, the white dwarfs are characterized by exceedingly high values for the mean density"

To answer these questions it is convenient to start by imagining what the high pressure inside the white dwarf depends on. To this respect Chandrasehkar (1939: 412) claims that "The clue to the understanding of the structure of these stars was discovered by R. H. Fowler[5], who pointed out that the electron gas in the interior of the white dwarfs must be highly degenerate … the white dwarfs can, in fact, be idealized to a high degree of approximation as completely degenerate configurations."

Indeed, if inside the white dwarfs there were a massive presence of hydrogen, the pressure at its centre and the very high temperatures inside it would bring about thermonuclear reactions that would produce luminosities much greater than that emitted by white dwarfs. As this is not the case, it is obvious that such reactions do not take place inside them. As Hansen et al. (2004: 474) claim "the total mass of hydrogen cannot much exceed $10^{-4}M_{\odot}$ because, if it did, nuclear burning would occur. Similarly, the helium layer mass should not exceed about $10^{-2}M_{\odot}$."

Fermions are particles subject to Pauli's Exclusion Principle which prohibits two particles from sharing the same quantum state. If the gas were a gas of electrons, which are fermions, as its temperature decreased, more and more electrons would tend to occupy the lowest energy levels. Since obviously, not all electrons can be at the fundamental level, once the fundamental states are occupied the electrons would continue to occupy the states with the lowest energy levels. At $T = 0\ K$ only the lower energy levels would be occupied. An electron gas with these characteristics is said to be *completely degenerate*. The maximum level of energy that can be reached by the electrons of a completely degenerate gas is called *Fermi energy*.

If we suppose an ideal electron gas, where the electrons move with equal momentum p and interact with each other through perfectly elastic collisions, then, turning to atomic physics, to the Heisenberg indeterminacy principle of quantum physics, to Pauli's Exclusion Principle, to the concept of electron degeneracy, etc., we obtain, in the relativistic limit $(\mathrm{v} \approx c)$, the pressure caused by the electron degeneracy. Equating the values of the electron degeneracy pressure and of the central pressure inside the white dwarf, we obtain *Chandrasekhar's mass limit* formula, M_{Ch}, (Ostlie

[5] The American physicist Ralf Howard Fowler (1888–1944) made in 1926 "the fundamental discovery that the electron assembly in the white dwarfs must be degenerate in the sense of the Fermi-Dirac statistics." (Chandrasehkar 1939: 451, Bibliographical Notes 1). And Shapiro and Teukolsky (2004: 56) claim: "In December 1926, R. H. Fowler, in a pioneering paper on compact stars, applied Fermi-Dirac statistics to explain the puzzling nature of white dwarfs: he identified the pressure holding up the stars from gravitational collapse with *electron degeneracy pressure.*"

On the history of the theory of white dwarfs see also Shapiro and Teukolsky (2004: 55–56).

and Carroll 1996: 590) that gives – in solar masses – the maximum mass value, 1.4 solar masses, which a white dwarf can support. This is the most important formula of the white dwarf theoretical model. This model offers the best theoretical explanation available so far of white dwarfs.

4 Conclusion

The use of a triple hypothesis: a completely degenerate electron gas with equal electron momentum in the relativistic limit, has been shown as a very fertile procedure, and the *combination* of Classical Mechanics, Quantum Mechanics and Relativity Theory – on the other hand, incompatible with each other – has allowed to build a model that provides the theoretical explanation of the white dwarfs *natural kind*.

In addition, the theoretical value of M_{Ch} is verified empirically. No observed white dwarf has ever exceeded the Chandrasekhar mass limit value. This justifies the claim that the white dwarf model offers the *best explanation* of the observations provided by this kind of stars. But as it is a theoretical model built *preductively* combining the aforementioned theories we can state without any doubt that the theoretical model of white dwarf is the result of a successful *sophisticated abduction*.

But what about causation? The construction of the theoretical white dwarf model is a complex process, far removed from a *proximate explanation* of a physical event. The *causes* that supposedly contribute to the state of *white dwarf* are *distributed* or *disseminated* among the principles and situations that allow the preductive construction of the theoretical model: Pauli Exclusion Principle, completely degenerate ideal electron gas, Fermi Energy, Heisenberg Uncertainty Principle, uniform distribution of electron moments, relativistic velocities of the electrons, etc. These assumptions build a *network of 'causal' hypotheses* necessary for the construction of the model; none of them is dispensable.

To summarize: the sophisticated abduction that provides a reasonably good theoretical explanation of the white dwarfs implies a dissemination or distribution of the causation in a multiple-cause network. The question of the search for truth, the whole truth, and nothing but the truth of the complex phenomenon investigated neither arises, nor is necessary.

Acknowledgment. I am very grateful to two anonymous referees for their valuable comments on this article.

References

Aristotle (1975) Posterior analytics. Barnes J (eds). Clarendon Press, Oxford

Berzelius JJ (1812/1813) Essay on the cause of chemical proportions, and on some circumstances relating to them: together with a short and easy way of expressing them. Ann. Philos 2, 3

Bessel FW (1844) On the variations of the proper motions of *Procyon* and *Sirius*. Mon Not R Astron Soc 6:136–141

Chandrasehkar S (1939) An introduction to the study of stellar structure. Dover Publications, New York

Ferrel W (1856) An essay on the winds and the currents of the oceans. Nashville J Med Surg
Galilei G (1954) Dialogues concerning two new sciences. Dover Publications, New York (trans: Crew H, de Salvio A)
Hadley G (1683) Concerning the cause of the general trade-winds. Philos Trans (1683–1775) 39:58–62 (1735–1736)
Halley E (1686) An historical account of the trade winds, and monsoons, observable in the seas between and near the tropicks, with an attempt to assign the phisical cause of the said winds. Philos Trans (1683–1775) 16:153–168 (1686–1692)
Hammer F (ed) (1963) Johannes Kepler gesammelte werke. Band VIII: Mysterium Cosmographicum, De Cometis, Hyperaspistes. C. H. Bleck'sche Verlagsbuchhandlung, München
Hansen CJ, Kawaler SD, Trimble V (2004) Stellar interiors. Physical principles, structure and evolution, 2nd edn. Springer, New York
Harman G (1965) The inference to the best explanation. Philos Rev 74(1):88–95
Kepler J (1992) New astronomy. University Press, Cambridge (trans: Donahue WH)
Koyré A (1961) La révolution astronomique. Copernic, Kepler, Borelli. Hermann, Paris
Lewis D (1986) Causal explanation. In: Lewis D (ed) Philosophical Papers, vol II. University Press, Oxford
Lipton P (2001) What Good is an Explanation? In: Hon G, Rakover S (eds) Explanation, theoretical approaches and applications. Kluwer, Drodrecht
Magnani L (2001) Abduction, reason and science. Processes of discovery and explanation. Kluwer Academic/Plenum Publishers, New York
Magnani L, Belli E (2007) Abduction, fallacies and rationality in agent-based reasoning. In: Pombo O, Gerner A (eds) Abduction and the Process of Scientific Discovery. Colecçao Documenta, Centro de Filosofía das Ciências da Universidade de Lisboa, Lisboa, 283–302
Newton I (1782) Opera quae exstant omnia. Tom IV. London. Samuel Horsley, ed. Facsimile edition. Friedrich Frommann Verlag, Stuttgart
Ostlie DA, Bradley WC (1996) Modern stellar astrophysics. Addison-Wesley Publishing Co., Inc., Reading
Peirce ChS (1965) Collected papers, CP. Harvard University Press, Cambridge
Rivadulla A (2008) Discovery Practices in Natural Sciences: From Analogy to Preduction. Revista de Filosofía Núm. 33(1):117–137
Rivadulla A (2015) Abduction in observational and in theoretical sciences. Some examples of IBE in palaeontology and in cosmology. Revista de Filosofía 40(2):143–152
Rivadulla A (2016) Complementing standard abduction. Anticipative approaches to creativity and explanation in the methodology of natural sciences. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology. Logical, epistemological and cognitive isues. SAPERE, vol 27, Springer, Cham, 319–328
Rivadulla A (2018) Abduction, Bayesianism and best explanation in physics. Culturas Científicas 1(1):63–75
Rivadulla A (2019) Causal explanations: are they possible in physics? In: Matthews MR (ed) Mario Bunge Centenary Festschrift. Springer
Rosen E (1959) Three copernican treatises. 2nd (edn), revised. Dover Publications, New York
Shapiro SL, Teukolsky SA (2004). Black holes, white dwarfs, and neutron stars. The physics of compact objects. Wiley-Vch Verlag GmbH, Weinheim
Whewell W (1847) The philosophy of the inductive sciences. Part one and part two, 2nd edn. Frank Cass and Co. Ltd, London

Defining a General Structure of Four Inferential Processes by Means of Four Pairs of Choices Concerning Two Basic Dichotomies

Antonino Drago(✉)

Naples University "Federico II", I, Naples, Italy
drago@unina.it

Abstract. In previous papers I have characterized four ways of reasoning in Peirce's philosophy, and four ways of reasoning in Computability Theory. I have established their correspondence on the basis of the four pairs of choices regarding two dichotomies, respectively the dichotomy between two kinds of Mathematics and the dichotomy between two kinds of Logic. In the present paper I introduce four principles of reasoning in theoretical Physics and I interpret also them by means of the four pairs of choices regarding the above two dichotomies. I show that there exists a meaningful correspondence among the previous three fourfold sets of elements. This convergence of the characteristic ways of reasoning within three very different fields of research - Peirce's philosophy, Computability theory and physical theories - suggests that there exists a general-purpose structure of four ways of reasoning. This structure is recognized as applied by Mendeleev when he built his periodic table. Moreover, it is shown that a chemist applies all the above ways of reasoning at the same time. Peirce's professional practice as a chemist applying at the same time this variety of reasoning explains his stubborn research into the variety of the possible inferences.

.

Keywords: Dichotomy on the kind of mathematics · Dichotomy on the kind of logic · Peirce's four ways of reasoning of computability theory · Four prime physical principles · General structure of ways of reasoning · Mendeleev's ways of reasoning · Chemical origin of peirce's reasoning

1 Introduction

The aim of this paper is to define four main ways of reasoning (in the following: WoRs; in particular, inductive and abductive reasoning).

In the first part of the present paper I will summarize what I have shown in previous papers: (*i*) Peirce's writings on the inference process of abduction may be best interpreted by means of intuitionist Logic. (*ii*) Beyond the declared inference processes - deduction, induction, abduction - Peirce's writings on both the criticisms of Descartes' philosophy and the characterization of the main logical features of a computer implicitly made use of a fourth inference process, which I called limitation (Drago 2013).

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 298–317, 2019.
https://doi.org/10.1007/978-3-030-32722-4_18

I will examine WoRs through the most basic notions possible. Recent research on the foundations of both Mathematics and Logic has suggested two dichotomies; one regarding the two formal kinds of Mathematics - either classical or constructive-, corresponding to the two kinds of the philosophical notion of infinity; another one on the two formal kinds of mathematical Logic - either classical or intuitionist-, corresponding to the two kinds of the philosophical notion of the organization of a theory. These dichotomies are traced back to Leibniz's two labyrinths, which the human mind encounters in its reasoning (Leibniz 1710). They constitute the foundations of science. (Drago 1994; Drago 2012) By means of the four pairs of choices regarding the above two dichotomies I have characterizes as well-defined logical processes both Peirce's four inferential processes and the four WoRs of Computability Theory (=CT) - i.e. recursion, minimalization, oracle and undecidabilities. Moreover, these two sets of fourfold WoRs prove to be in a mutual correspondence; in particular, the inference process of an abduction corresponds to CT's WoR of an oracle; as such, it is accurately defined in both mathematical and logical terms. (Drago 2007; Drago 2016).

In the following two parts of present paper I will obtain three main results. (*i*) Recently, I have recognized within theoretical Physics (classical Chemistry included) four prime principles of reasoning, i.e., causality, extremants, physical existence of a mathematical object, impossibilities, all characterized by means of the four pairs of choices regarding the two dichotomies. (Drago 2015) There exists a semantic correspondence among these prime physical principles of reasoning and the two fourfold WoRs of both CT and Peirce's philosophy. In particular, in a physical theory an abduction corresponds to the claim to attribute physical existence to a mathematical object (e.g. in geometrical optics, to claim that a straight line represents a light beam). (*ii*) All of this is evidence of the same structure of four WoRs which is common to the three fields of research, i.e. Peirce's philosophy, CT and theoretical Physics. (*iii*) The various well-founded physical theories – built up over the centuries - enjoy this structure of reasoning and this constitutes evidence for its adequately representing the scientific WoRs about the real world. This structure was substantially reiterated by CT and was anticipated by Peirce's philosophical reflection. Hence, this structure represents not only the WoRs of a variety of formal scientific theories, but also a philosophical conception. This convergence on the four WoRs, obtained from three very different fields of research, constitutes sufficient evidence for its both philosophical and logical completeness. (*iv*) I will recognize Peirce's and CT's WoRs in Mendeleev's reasoning when he built the Table of elements, in particular the abduction inference. *v*) Yet, CT differs from a common physical theory, which argues mainly through a single prime principle of reasoning, because it argues by means of all WoRs at the same time. (*vi*) A classical chemist also reasons in the latter way. This fact provides an interpretation of the great work performed by the professional chemist Peirce in discovering all possible WoRs, as well as his insistence on the notion of abduction, really an essential inference process of Physical-Chemistry, but completely ignored by theoretical physicists, owing to its elusive nature, which will be explained later.

2 Two Dichotomies as the Foundations of Physical Theories

I exploit two decisive results obtained by the investigations into the foundations of science. Half a century ago two basic philosophical notions received clear-cut formal definitions. The notion of infinity, in which philosophers had distinguished actual infinity (AI) and potential infinity (Pl), has been formalized as two well-defined formal systems; on one hand, traditional classical mathematics, which since the 17th century has relied upon Al (e.g. through notions as infinitesimals and Zermelo's axiom); and, on the other, constructive mathematics, relying on (almost) only PI (Markov 1962; Bishop 1967).

A more laborious historical process led to a formal definition of the philosophical organization of a theory. Aristotle suggested that a scientific theory has to be organised through a pyramidal system of deductions drawn from a small number of axioms. Of course, this organization is governed by classical logic. For a very long time the mainstream maintained that classical logic is unique. Eventually, in last Century the logicians recognized a plurality of kinds of logic, all formalized in mathematical terms; in particular, intuitionist logic was recognized as being on a par with classical logic. Moreover, by means of a comparative analysis of some theories - pertaining to Logic, Mathematics and Physics – which exhibit a different organization from the deductive one, e.g. classical Chemistry, I discovered that each of these theories makes use of propositions of a particular kind. They are doubly negated propositions which are not equivalent to the corresponding affirmative propositions owing to the lack of evidence for the contents of the latter ones (DNPs).[1]

An instance of a DNP in theoretical Physics is the following one: "Motion without end is impossible". (Dugas 1950, p. 121) In such a case the double negation law fails since this proposition is not equivalent to: "Every motion has an end", which, as a

[1] Notice that a single word, in particular a modal word, may be equivalent to a DNP; e.g. possible = it is not the case that it is not" (this kind of word will be underlined with dots). More in general, it is well-known that modal logic may be translated by means of its S4 model into intuitionist logic. (Chellas 1980, 76ff.) Notice that the current usage of the English language exorcises DNPs as pertaining to primitive languages. Moreover, some linguists maintain that those who speak by means of DNPs want to be, for instance, unclear. (Horn 2002, pp. 79ff.; Horn 2010, pp. 111–112) On the contrary, it is easy to show that the DNPs pertain to scientific research in Logic, Mathematics, Physics and classical Chemistry. In Logic the translation from classical logic to intuitionist logic is performed by doubly negating the propositions of the former logic. (Troelstra and van Dalen 1988, p. 56ff.) In Mathematics it is usual to develop a theory in order to make it "without contradictions" (here and in the following I underline the negative words belonging to a DNP for an easy inspection by the reader); owing to Goedel's theorems, it is impossible to state the corresponding affirmative proposition. i.e. the consistency of the theory at issue. In Mathematics and in theoretical Physics it is usual to study in-variant magnitudes; this adjective does not mean that the magnitudes remain fixed. Moreover, substantial advances were achieved in Mechanics by means of the above mentioned, methodological principle of the impossibility of motion without end. In Chemistry, in order to solve the problem of what the elements of matter are, Lavoisier defined these unknown entities by means of a DNP: "If we link to the name of elements… the idea of last term arrived at by [chemical] analysis, all the substances which we were not able to decompose by any means are for us elements: (Lavoisier 1862–92, p. 7) where the word 'decompose' carries a negative meaning since it stands for 'non-ultimate' or 'non-simple'.

scientific law, is false; because nobody is capable of operatively determining, owing to the a priori unknown friction function, the final point of say the Earth's trajectory, or also of a ball struck with a cue on a billiard table before this end occurs. In the last century the scholars of mathematical logic achieved a crucial result; i.e. the validity or not of the double negation law represents the best discriminating mark between classical logic and most non-classical logics, above all intuitionist logic (Prawitz and Melmnaas 1968; Grize 1970, pp. 206–210; Dummett 1977, pp. l7–26; Troelstra and van Dalen 1988, pp. 56ff.). This failure of the double negation law qualifies the former proposition as belonging to non-classical logic, in particular, intuitionist logic.[2]

This logic governs a different model of organization of a scientific theory, which I have obtained by means of a comparative analysis of all past scientific theories which present an organization other than a deductive one; in particular, Lobachevsky's theory of non-Euclidean geometry (Lobachevsky 1955). Each of these theories is aimed at solving a basic problem by inventing a new scientific method by means of *ad absurdum* arguments. I called this model of organization a problem-based organization (PO), whereas I called AO the Aristotelian organization of a deductive kind. In such a way the philosophical notion of two kinds of organization of a theory is translated into a formal dichotomy between the two main kinds of mathematical logic. In sum, we have, on the one hand, classical logic, governing AO theories (e.g. Euclid's *Elements*) and, on the other, intuitionist logic, governing PO theories (Drago 2012).

The following six kinds of analysis have corroborated these two dichotomies as the foundations of science:

(i) A clear recognition of the foundations of Newton's mechanics as constituted by the following two choices; the deductive organization, starting from his celebrated three principles, AO, and the use of an idealistic mathematics, i.e. infinitesimal analysis, hence the choice IA. (Drago 1988) *(ii)* The rational re-construction of Lazare Carnot's mechanics, which completed Leibniz's effort to suggest an alternative theory to Newton's mechanics; (Drago Manno 1989; Drago 2004) it is based on the problem of the impact of bodies (PO) and its mathematics is plain algebraic-trigonometric mathematics; hence, its two choices regarding the two dichotomies diverge from those of Newton's. *(iii)* The rational re-construction of Sadi Carnot's thermodynamics, which was the first alternative physical theory to Newton's mechanics (Drago and Pisano 2000); it manifests the alternative choices to Newton's, makes use of elementary mathematics and is based on the problem of the highest efficiency in the conversion of heat into work. *(iv)* The interpretation of the large number of the new theories developed at the time of French revolution; they differ from each other in the pairs of choices. (Drago 1989) *(v)* The interpretation of the revolutionary role played by Einstein's first paper on quanta, as manifesting a complete alternative to Newton's foundations, an alternative that can be traced back to the difference between the fundamental choices of this theory and those of Newton's (Drago 2013). (*vi*) A systematic interpretation of all categories applied by the historians of Physics, in particular

[2] As a matter of fact, Grzegorczyk (1964) independently proved that the production of new results by science experimental may be formalized through propositions belonging to intuitionist logic, that is, a logic using DNPs.

Koyré's and Kuhn's categories which translate the pairs of choices - AI&AO of Newton's mechanics - into subjective terms (Drago 2017).

Viceversa, each pair of choices determines one out four models of a scientific theory (MSTs). I baptized the MST of the choices AI&AO, upon which Newton's mechanics relies, the Newtonian MST. Instead, Classical Chemistry, L. Carnot's Mechanics, S. Carnot's Thermodynamics, Lobachevsky's non-Euclidean Geometry, Einstein's first theory of quanta,[3] etc. (Drago 1996) all belong to the Carnotian MST, whose choices are PI&PO; whereas Descartes' theory of geometrical optics is representative of the Descartesian MST of the choices PI&AO; Lagrange's theory of mechanics is representative of the Lagrangian MST whose choices are AI&PO.

Notice that the two dichotomies are more powerful categories than any category suggested by previous philosophers of science, most of whom suggested a single notion (i.e. causality, determinism, economy of thinking, extremants, probability, etc.); the dichotomies are instead two independent notions. Moreover, they are two very particular notions, i.e. dichotomies; as such, they allow four choices; hence, instead of a monist or at most a twofold scheme, a fourfold scheme constitutes the foundations of science. In addition, previous scholars looked for either philosophical, informal notions (e.g., space, time, set, determinism, causality, etc.) or formal notions (ruler and compass, infinitesimals, Euclidean geometry, calculus, Newton's mechanics, etc.) as the foundations of science. Instead, the above dichotomies constitute, at the same time, philosophical notions (infinity, organization) and formal scientific notions (or even theories); indeed, each dichotomy is formalized in mathematical terms. Hence, this double faced nature allows their application to fields of reality in both formal and informal terms.

I add that these dichotomies can be traced back to a noble philosophical father, Leibniz. He stressed "two labyrinths of human mind": (1) the notion of infinity: either actual infinity or potential infinity; (2) "either law or freedom". (Leibniz 1710, Preface) He was unable to decide whether each labyrinth is solvable or not. Subsequently, no one has resolved them by scientific means. This fact suggests that each labyrinth is actually a dichotomy for human reason. Of course, the first of Leibniz's labyrinths concerns the same above dichotomy of the two kinds of mathematical infinity. The second labyrinth corresponds to the above dichotomy of the two kinds of organization, provided that this organization is considered from a subjective viewpoint: either obedience to a compulsory law derived from fixed principles, or the freedom to creatively discover a method for solving a given problem.

[3] Einstein qualified his paper, suggesting the physical existence of light quanta (Einstein 1905a), as his "most revolutionary paper". (Einstein 1905b) It explicitly presents the dichotomy of infinity in mathematics, and implicitly, yet in an almost rigorous way, presents the dichotomy in both the kind of organization and the kind of logic. Remarkably, it may be considered the most revolutionary paper in general, because it was the only paper to present the two dichotomies. (Drago 2013) Instead L. Carnot (1803) and Lobachevsky (1955) obtained the same result but through books, respectively the former a book founding an alternative Mechanics to Newton's and the latter a book on non-Euclidean geometry.

3 Improving Peirce's Philosophical Characterization of Both the Behaviour and WoR of a Computer

Among past philosophers Peirce was the only one with a background as a chemist and moreover one who worked as a scientist (he was mainly a geophysicist). He also was one of the few philosophers that developed their philosophical systems on the basis of scientific research. His philosophy was baptized by him as pragmatism, whose meaning *grosso modo* corresponds to the method of experimental science. In fact, his philosophical reflection was primarily concerned with the methods of inquiry and the growth of knowledge. In particular, Peirce was one of the first philosophers to ponder on the "thinking machines" and among these philosophers he was certainly the most intelligent. Being a pragmatic philosopher, Peirce basically referred his thinking to operative processes; current computability theory (CT) also refers to an operative process (of calculation). Moreover, his reasoning was mainly aimed at solving problems, as CT also does. Furthermore, he made a great contribution to determining how to conceive CT. In a more specific terms, Peirce stressed the "impotencies" of such "thinking machines" (Peirce 1887), pp. 168–169).

His primary interest in investigating scientific research was to characterize the WoRs of man's mind. Of course, he studied deduction and induction, adding a new inferential process, "abduction" (Fann 1970, p. 26), by which he meant mainly "the reasoning by which we are led to adopt a hypothesis" (2.102). Moreover, I have suggested that Peirce introduced, although he was unaware of it, a fourth process of reasoning, which, being of a limitative kind, I called "limitation". Three of Peirce's writings on a crucial philosophical subject - his criticism of Descartes' basic tenets. (Peirce 1868b; Peirce 1896b; Peirce 1869) – actually makes use of a fourth kind of reasoning, establishing the "incapacities" of human reasoning. Moreover, some of Peirce's reflections upon computers stressed their "impotencies" (Peirce 1887, pp. 168–169). On the basis of his writings on both Descartes' tenets and computers' impotencies (Drago 2014).

Although Peirce's presentation of these WoRs is disputable because his writings did not even accurately define the two inferential processes of induction and abduction. I conclude that Peirce actually substantially suggested four WoRs: Deduction, induction, abduction and limitation (including both "incapacities" and "impotencies"). This framework is wider than the usual one, often reduced – as Peirce himself lamented (8.384) - to the deductive WoR only; or, at most, it is commonly enlarged to include elements of induction; while abduction is commonly ignored; at worst, any limitative reasoning is considered to be a useless constraint on the scientific research.

4 A Semantic Correspondence Between Peirce's Four Inferential Processes and CT's Four Mathematical WoRs

Of course, Peirce's framework of the four inferential processes is of a philosophical nature. In order to move towards a formal characterization of it, let us analyze his fourfold system from a new viewpoint; which is expressed by means of a formal language developed over millennia, i.e. mathematics.

Through this language CT has suggested four distinct techniques of calculation which the mind performs as processes of reasoning; i.e. recursion. minimalization. oracle[4] and undecidabilities. Let us now compare Peirce's four inference processes with the four mathematical processes characterizing CT.

Notice that this comparison concerns, on one hand, formal WoRs of a mathematical or logical nature and, on the other, informal WoRs of a philosophical nature. The comparison will test whether a formal WoR is an instantiation of a more general WoR which is defined in philosophical terms; hence the correspondence to be established cannot be anything more than an equivalence of a semantic nature.[5]

It is easy to see a correspondence between Peirce's two inference processes, i.e. deduction and limitation, and two particular CT WoRs. Indeed, CT's *recursion* represents a particular instance of deductive reasoning from the formula of the recursive function playing the role of an axiom, from which the n-th result is obtained by the n-th iteration of the same deduction process. Actually, Peirce (Peirce 1881) was one of the first mathematicians to suggest the mathematical definition of recursion as a specific instance of a deductive WoR.

Moreover, CT's *undecidabilities* at a glance appear to be a formalization through exact mathematical tools of Peirce's notion of computers' impotencies.

About Peirce's two remaining inferential processes, i.e. induction and abduction, we have to take into account that he never accurately defined them. (Fann 1970, pp. 6, 9–10, 31) Thus in order to obtain accurate definitions of them I take advantage of the formal characterizations of CT's two remaining mathematical WoRs, i.e. minimalization and oracle. Notice that the justification of a minimalization is given by a mathematical calculation generating it; whereas the justification of abduction is given by an *a posteriori* verification of a logical nature.[6] Thus, I suggest that the Peirce's two

[4] It is roughly defined as a black box which is able to decide certain decision-making problems, otherwise unsolvable, through a single operation. It corresponds to the algebraic procedure of transcendental extension. (Odifreddi 1989, p. 175) It is more precisely defined as follows: "A number m that is replaced by $G(m)$ in the course of a G-computation… is called an oracle to query to the G-computation" (where a G-computation is a computation of a partial recursive function G under a specific condition referring to total functions. (Davis et al. 1995), pp. 197ff.).

[5] CT accustomed scientists to comparing an informal notion of computability with formal notions (recursion, Diophantine equations, λ-calculus, etc.). It is proved that all the formal notions of computability are equivalent and "hence" (Turing-Church's thesis) they may be equated to the informal notion. In our case the arguments will be looser than those of CT, because they are aimed at obtaining not mathematical results, but semantic results concerning different representations of WoRs.

[6] E.g. it is an abduction that suggests the number $\sqrt{2}^{\sqrt{2}}$ for solving the problem whether there exist two irrational numbers a and b such that a^b is a rational number. Proof: either $\sqrt{2}^{\sqrt{2}}$ is the desired rational number, or $\sqrt{2}$ elevated to $\sqrt{2}^{\sqrt{2}}$ solves the problem. One more instance is Lobachevsky' suggestion of a definition for a parallel line as that line that with the least displacement crosses the basic line; (Lobachevsky 1955, prop. 16) this definition is then justified by logical means, i.e. two theorems which make plausible this definition. In both cases the validity of the solution is verified by a logical argument.

inference processes are mutually distinguished according to their kinds of justification, respectively an *a priori* mathematical one and an *a posteriori* logical justification.[7]

After these qualifications of induction and abduction, Peirce's inference of induction may be accurately defined as obtained by means of a specific mathematical process (continuity, infinitesimals, limits, extremants, involution, etc.), which are all included by CT's mathematical WoR which produces a *minimalization* (or *maximalization*); whereas Peirce's inference of abduction may be defined as an instantiation of a CT's computing process obtaining from *an oracle* an answer suggesting an element that is a posteriori justified in logical terms, i.e. by its not contradicting its original mathematical context (Drago 2014; Drago 2016).

Let us stress the elusive nature of abduction. It is a common opinion among scientific researchers that when a result is known as possible, because it has been already obtained by others, it is just a matter of time before the same result is obtained again. That means also that once the result of abduction is logically justified since it is shown to work, it is a matter of (a short) time before it is discovered that either an inductive or a deductive process obtain the same result. Of course, the latter two methods of discovery are considered more cogent than an abduction, whose purely logical proof may be open to metaphysical notions and considerations. For this reason, once a result is obtained by means of abduction, scientists promptly replace it with a more "respectable" inference. This explanation of the elusive nature of abduction holds true even more in the case of a physical theory, where a logical justification is commonly considered to be too abstract with respect to experimental reality; as a matter of fact, in the history of theoretical physics a physicist has never claimed a result by an argument based on abduction, if not as a mere guess, motivating the search for either a mathematical calculation, or a theoretical deduction, or an eminently experimental datum, whose evidence provides the correct justification of the result. It is for this reason that almost all scientists have avoided presenting abduction with impunity. This custom has excluded abduction from the commonly recognized experience of scientific reasoning belonging to the most important area of scientific inquiry.

Although the previous comparison of intuitive philosophical ideas, i.e. Peirce's definitions of inferences, and formally defined mathematical ideas - i.e. CT' mathematical processes-, allow only philosophical considerations, some considerations of this kind seem important: (*i*) Peirce's philosophical effort to qualify the potentialities of "thinking machines" not only anticipated better than anyone of his time a philosophy of CT, but also the inference process of limitation, and hence all kinds of inference processes. (*ii*) Rather than a metaphysical basis - which Newton chose to give to his Mechanics (see his metaphysical notions of absolute space, absolute time and force-cause)-, or the basis of an empiricist philosophy - given by Lazare Carnot to his Mechanics (all its notions and also principles are of an empirical nature; Drago 2004) -, a pragmatist philosophy is CT's philosophical basis. (*iii*) This philosophy was formulated by the scientist-philosopher Peirce half a century before CT's birth; hence he

[7] Later, Peirce (1958, vol. 8 p. 58) called induction and abduction respectively "Quantitative and Qualitative induction". (2.755; 6.526) As usual among Peirce's scholars, a reference to (Peirce 1958) it will be given by a first number denoting the volume and a second number denoting the issue.

has to be considered the philosophical father of CT. (*iv*) This philosophical basis of CT does not concern basic notions or principles – as in both Newton's Mechanics and L. Carnot's Mechanics - but that which is the main subject of CT, i.e. WoRs – which in science constitutes a higher level of conceptualization than notions and principles. (*v*) The above illustrated correspondences provide Peirce's philosophical system of WoRs, abduction included, with exact definitions. (*vi*) The correspondence of Peirce's philosophical inference processes with scientific experience of CT's WoRs along almost a century suggests a reasoning structure which is simultaneously informal and formal in nature.

From this correspondence one *may* suggest that Peirce's four inferential processes represent all possible inferential processes; and *viceversa*, that for philosophical reasons CT's four WoRs *may* represent a complete framework of formal WoRs. In the following Sect. 6 we will add decisive evidence for supporting these theses.

5 Recognition of Peirce's Inference Processes in Mendeleev's WoRs Aimed at Formulating His Table of Elements

In the following we will recognize the previous system of four WoRs as the set of WoRs which Mendeleev made use of when formulating his table of the elements of matter. In the history of science this case-study is unique because Mendeleev not only reasoned in a variety of ways in order to obtain his result, but he also described his WoRs.[8] An important publication by Mendeleev (one of Faraday's Lectures, quoted by Scerri 2007, pp. 109–110) illustrates the method that Mendeleev exploited in order to construct his periodic table through physical and chemical experimental data. The specific WoR actually referred to by Mendeleev is in square brackets.

1. The elements, if organized according to [the growing numbers of the] atomic weights [*deduction-recursion*], show an evident periodicity [*limitation*] of the properties.
2. Elements that are similar in their chemical properties have atomic weights that are either nearly equal (e.g., Platinum, Iridium, Osmium, etc.) [the similarity of elements is regularly represented by the atomic weight as well as all other properties, which means a contiguity of the values of each of their parameters; *deduction-recursion*] or [in the case of the same valence, in the sense of a variable with a *limited* range] that increase regularly [*recursion-deduction*] (e.g., Potassium, Rubidium, Cesium).

[8] At MBR '18 conference A. Rivadulla presented a new instance of abduction, i.e. Bessel's discovering of the double nature of the star Syrius. During the discussion after his presentation I have remarked that Bessel's words "non unfitting" constitute a DNP ("It is not non-usual that…") which is logically equivalent to Peirce's word "suspect". It would interesting to re-visit all the cases of discovery of a new planet - they surely constituted instances of abductions - under the light of Peirce' three statements.

3. The organization of the elements, or groups of elements, in the order of their atomic weights, corresponds to their so-called valences [*limitation*], as well, to some extent, to their characteristic chemical properties - as is clear in another series [*deduction-recursion*] - in that of Lithium, Beryllium, Boron, Carbon, Nitrogen, Oxygen and Iron.
4. The elements that are most widespread [in nature] have small atomic numbers [an experimental fact of geology, not chemistry].
5. The value of atomic weight determines [*deduction*] the character of the elements, just as the value of the molecule determines the character of a compound.
6. We must expect [owing to an *ad absurdum* argument, supporting an *abduction of the decisive hypothesis of* the periodicity for constructing a theory conceived as a systematic table] the discovery of many elements that are still unknown [*abduction*], for example elements similar to Aluminium and Silicon, the atomic weight of which [*induction*] should be between 65 and 71.
7. The atomic weight of an element can sometimes be corrected by the knowledge of the [atomic weights of the] adjoining elements [*induction*]. Therefore the atomic weight of Tellurium must be between 123 and 126, it cannot be 128.
8. Certain characteristic properties of the elements can be predicted by their atomic weights [*induction*].

Let us now interpret in more detail these reflections of Mendeleev's through the above four WoRs (according to *grosso modo* the order of Mendeleev's illustration).

Recursion-Deduction. In chemistry it corresponds to Prout's hypothesis: each element can be obtained as a multiple of a same element (Hydrogen). Therefore, the elements can be listed as a series of growing values of a parameter, e.g. the atomic weight. However, in Mendeleev's table the *atomic weight* expresses this growth in an irregular way;[9] years later, the *atomic number* parameter and then the *number of electrons* will give the exact recursion (albeit the most elementary kind of recursion). In addition, the same kind of inference as before is applied to the elements sharing the same valence, i.e. belonging to the same chemical group. This WoR occurs at points 1 (first part), 2 (first and third part), 3 (second part), 5.

Induction Through a Limit Process on (Rational[10]*, Hence Constructive) Numbers*. This case occurs when we consider the series of the experimental determinations of the values of a parameter – e.g. the atomic weight - of the surrounding group of a given or supposed element *X*. From these experimental results, we can obtain the value for the element *X* through a limit process (actually, by means of a limit process including very few approximating values). The series of values are considered as being no more than an approximation to the desired value. Through the use of this kind of inference each

[9] Peirce suggested rendering uniform the increments of the atomic weights by supposing that they were inaccurate for several reasons; hence he added to the atomic weight of each element a weight of up to plus or minus 2,5.

[10] Let us recall that the result of each measurement is only a rational number - because it is represented by a series of decimal digits truncated to the best approximation.

element is characterized by a list of approximatively determined values of its chemical-physical properties. All Mendeleev's *analogies*, formulated as "averages" on triads or octaves of elements, are such limit processes which represent inductions. This WoR occurs at points 6 (last part), 7 and 8.

Limitation. Valence (of course, among the different valences enjoyed by an element we will consider the basic one). Its nature is of a limitation WoR because (1) it confines a variable to a finite interval; hence, by analogy with geometry, when the radius of curvature of the space is finite, the geometry is elliptic and its lines are periodic in nature; more precisely, valence defines the constraints within which a chemical element can combine itself with the other elements; or defines the constraints within which a chemical group is located; (2) it is defined by a DNP: it is not possible to consider as homologues two elements with different valences. It is by reflecting on the similarity of elements with the same valence that Mendeleev made "the crucial discovery" of periodicity [17, p. 105, p. 119]; which he then combined with the previous recursive progression; it is precisely in this way that he obtained his MT. This WoR occurs at points 1 (second part), 2 (second part), 3 (first part).

Abduction as an oracle of a decisive hypothesis for constructing a theory ("the logic of [theory] pursuit", as Achinstein (1993) put it); this inference may be what I have called "Peirce's Principle". (Drago 2016). It concerns the table as a whole; it is of this nature the hypothesis that states the completeness of MT. By attributing importance to empty locations, an *ad absurdum* argument is implicitly stated: material reality would be absurd if it were to admit these voids in a series of material elements. This leads to the hypothesis that in the sequence of the elements it is impossible that, in that place, there does not exist a new element. To this proposition we apply a general principle of translation between two kinds of logic, i.e. the principle of sufficient reason, translating from intuitionist logic to classical logic.[11] We then infer from the relations with its neighbouring elements, that this new element must have similar characteristics to those possessed by them. This WoR occurs at points 6 (first part).

Table 1 summarizes both the suggested links between the four WoRs and Mendeleev's illustration of his reasoning for constructing his MT.[12].

In sum, we have obtained that Peirce's four inferential processes represent the WoRs which Mendeleev was aware of having applied when composing on a logical basis his systematic table. In particular, abduction plays a decisive WoR in his building the table.

[11] Peirce conceived abduction in a similar way to the application of the principle of sufficient reason: "Abduction "tries what *il lume naturale* … can do. It is really an appeal to instinct" (1.630) "Retroduction [read: abduction] goes upon the *hope* that there is sufficient affinity between the reasoner's mind and nature's to render guessing not altogether hopeless…" (1.121).

[12] The text of Faraday's Lecture (Mendeleev 1889) may be interpreted as concerning the same WoRs; recursion on atomic weights (p. 103); Limitation of the values of valence, "closed circle" (p. 104); induction (pp. 106-107); Abduction and induction (p. 117ff).

Table 1. Correspondences between Peirce's kinds of inferences and those suggested in Mendeleev's works

Mendeleev's WoRs	Peirce's inference processes	First physical principles
Series of the atomic weights of elements (Prout) (either in general or within a chemical group)	Recursion-deduction	Causality-deduction (Geometric optics, Newton's mechanics)
Atomic weight of an element obtained from a limit process performed on experimental data	Induction (idealistic or approximate)	Extremants
Valence that limits the groups of the elements. Periodicity of the properties of the elements	Impotence	Limitation (e.g., Impossibility of perpetual motion)
Completeness of the table	Abduction (as Peirce's Principle)	Principle of sufficient reason (Solution of the basic problem of a PO theory)
Best hypothesis of new element ("Analogy") through a mean of the properties of some neighbouring elements	Abduction (as an oracle of the best hypothesis)	Existence (ray of light in Geometric optics, fields in Electromagnetism)

6 Establishing Through the Four Pairs of Choices a Correspondence Among Peirce' Four Inference Processes, CT's Four WoRs and Four Prime Physical Principles

In order to improve this kind of analysis, let us consider one more scientific instantiation of the possible WoRs, i.e. those accumulated by a variety of physical theories over the last few some centuries. It is easy to recognize, in correspondence to the four representative theories of the four couples of choices regarding the two dichotomies -and hence the above listed four MSTs-, four prime principles of reasoning, respectively: causality, as embodied by the notion of force-cause of Newton's mechanics; extremants, as embodied by the principle of minimal action of Lagrange's mechanics; physical existence of a mathematical element, as embodied by a straight line which is considered a light beam of geometrical Optics; limitation, as embodied in Thermodynamics by the principle of the impossibility of perpetual motion (Drago 2011, Drago 2015).

As pertaining to a specific MST, each WoR is characterized by the pair of basic choices of its MST; these pairs are respectively: AI&AO, AI&PO, PI&AO, PI&PO.[13] Within theoretical physics each of these principles of rational WoR combines an

[13] The 18th Century saw the origin of a paradigm of considering not only Newton's mechanics, but theoretical physics as a whole as determined by the prime principle of force-cause. The considerable number of marvellous results thus obtained obscured the prime principles of all other theories, above all the prime thermodynamic principle of limitation together with the basic notion of entropy, which are at odds with the previous one. This paradigmatic view depreciated Thermodynamics as an immature, merely phenomenological, theory (see eg. Kuhn 1977). As a consequence, the paradigmatic view considered the so-called nomological deductive model of a scientific theory, which is derived from Newton's mechanics, to be the only one.

intuitive reality of reference, an operative method and a mathematical formalization (e.g. extremants).

Do the semantic contents of the four physical prime principles correspond to those of the four CT's WoRs? We will obtain an affirmative answer by characterizing also CT's Wors through the basic choices. First, let us investigate CT's WoRs of undecidability. Notice that only when AI is excluded can there be an indecidability; otherwise, a suitable ideal element decides any question (e.g. the idealistic Zermelo's axiom solves the constructively impossible problem of composing a new infinite set from an infinite number of infinite sets). Moreover, only in the context of a PO undecidabilities exist; otherwise, the desired decision is nothing other than a theorem derived from some a priori axioms (as within the AO theory of Euclidean geometry the following question: "Are equal two triangles if their three sides are equal?" is decided by a theorem derived from the well-known axioms). Hence, the WoR of undecidabilities is characterized by the pair of choices PI&PO, i.e. the same pair of choices determining the prime physical principle of limitation in theoretical physics of the Carnotian MST, in particular in S. Carnot's Thermodynamics. (Mach 1986, chp. XIX) It is easy to recognize that their semantic meanings are similar.

Recursion is eminently a tool of deductive reasoning: hence, it is determined by the choice AO. Moreover, at its birth (Goedel 1931) the mathematical technique of recursion was based on the choice PI; later, the notion of generalized recursion has been introduced; it manifestly appeals to ideal elements (e.g. the general recursive functions obtained by diagonalization processes on the totality of the primitive recursive functions; Davis et al. 1995, p. l05ff.). In sum, general recursion relies on the pair of choices AO&AI, i.e. the same pair of choices that in theoretical physics determines the (implicitly metaphysical) causal connection, as this pair does in Newton's mechanics.

Unbounded minimalization relies on AI, because its process of calculation appeals to a mathematical technique which in general is a non-constructive one. Moreover, it is aimed at solving a (computation) problem which is not solvable through ordinary means, hence it relies on the choice PO.

By resolving a decision-making problem without any evidence, an oracle represents a non-constructive element which plays only the role of an axiom, hence it introduces an AO theory; its context, in order to avoid metaphysical detours, has to be decidable, i.e. the choice of this theory is PI.

I conclude that the correspondence established by the four pairs of basic choices allows us to conclude that CT's four mathematical WoRs, are not only spontaneous inventions originating in a specific empirical context, but *constitute a foundational structure of CT*. The complexity of such a structure (including also an oracle, which is very different from the other WoRs) justifies why so many previous scholars were unsuccessful in recognizing it as a logical structure.

Moreover, the four foundational pairs of choices on the two basic dichotomies characterize a correspondence between CT's four mathematical WoRs *with* the four prime physical principles; so that the two quadruples are put in a one-to-one correspondence.

In previous papers I showed that the four pairs of basic choices regarding the two dichotomies also characterize Peirce's four processes of inference. (Drago 2008; Drago 2014) This result further qualifies the results of Peirce's philosophical investigation into inference processes (limitation included) as the anticipation of a well-defined philosophical system of WoRs; which represents "the logic of discovery", whose existence Peirce so often stressed (e.g. when he stated that "each chief step in science has been a lesson in logic"; 5.363). Notice that the actualization of such a logic denies Reichenbach's tenet: "Epistemology does not regard the process of thinking in their actual occurrence, this task is entirely left to psychology" (Reichenbach 1938, p. 5).

In sum, we have obtained that *the three fourfold WoRs*, which include both formal and informal aspects, by sharing the same *four pairs of basic choices*, may be put in a mutual, one-to-one correspondence.[14]

The following table summarizes all the above correspondences.

Table 2. Correspondence between CT's wors and the prime physical principles

Couples of basic choices	AI&AO	AI&PO	PI&AO	PO&PI
Model of scientific theory	Newtonian	Descartesian	Lagrangian	Carnotian
Prime principles in theoretical Physics	*Causality*	*Extremants*	*Physical existence of a mathematical object*	*Limitation*
WoRs in computability theory	*General recursion*	*Unbounded Minimalization*	*Oracle*	*Undecidabilities*
Peirce's four inferential processes	*Deduction*	*Induction*	*Abduction*	*Incapacities*
Nature of his inferential process	Conservative	Ampliative	Ampliative	Limitative
General logic principles	Principle of non-contradiction		Principle of sufficient reason	

Notice that previous results manifest a progression from philosophical notions, which is what Peirce's inference processes are, to prime physical principles, which combine formal and informal features, to CT's WoRs which are instantiated by mathematical processes.

[14] E.g. CT's oracle corresponds also in semantic terms to the prime principle attributing physical reality to a mathematical being; moreover, both are specifications - albeit within two different contexts - of the same philosophical notion of an abductive inference.

Their technical difference is an instance of the phenomenon of a radical variation of the (at least partially) formal representations of the philosophical notion of a WoR.

7 CT's Use of Four WoRs Diverges from that of Each Physical Theory

However, the comparison between CT and a physical theory (apart from Chemistry) indicates a surprising feature, whereas the theoretical development of a physical theory essentially makes use of only one reasoning principle, CT makes use of all mathematical WoRs at the same time. E.g. within CT the undecidabilities, already characterized as a PI&PO WoR, occur together with the general recursion functions, representing an AI&AO WoR.

This situation was perhaps caused by a rooted prejudice shared by the dominant group of mathematicians: being very useful for practical purposes, CT's theoretical content was negligible. In order to overcome such a prejudice CT's scholars have enlarged the theoretical import of their theory as far as possible in order to obtain the greatest possible number of theoretical results. This endeavour then led scholars to make free use of WoRs that are more powerful than, in particular, the WoR of the theory of primitive recursive functions: general recursion, unbounded minimalization, and oracle, all representing WoRs including idealized elements.

The application of different WoRs causes radical variations in the formal definitions of the various notions - e.g. either primitive or general recursion; either bounded or unbounded minimalization, either a computation step obtained by means of an oracle or a computation through an effective computation. But it causes also variations in meanings which are very subtle, owing to the imprecision of common language using informal notions whose variations in meaning may be misinterpreted as tolerable variations; so that they are not recognized as significant. For instance, surely few people notice the difference between "computable functions" and "computed functions", "computability theory" and "computations theory", etc.; in other words, between the meaning of a modal word which simply alludes to some content and the meaning of its objective name. Yet, each previous variation implies a variation of the kind of logic, either classical or modal logic, and hence, *via* S4 model (Chellas 1980, pp. 76ff.), intuitionist logic; when the different notions above are combined together, their two representative kinds of logic imply a contradiction in the use of the double negation law.

Notice that whereas a contradiction within an axiomatic system may be recognized after an indefinite number of steps of reasoning (this is the reason why one cannot prove directly the consistency of say Arithmetic), a contradiction within a theoretical system which combines WoRs - which are incompatible also in informal (verbal) aspects - is even more elusive. As a consequence, while an actually inconsistent theory may survive for a long time without scholars stumbling on a formal contradiction, a theory mixing different WoRs may survive for a long time before scholars accurately define a contradiction generated by the radical differences in both the WoRs and the meanings of their basic notions.

8 Chemists' Simultaneous Use of All WoRs

After the birth of classical Chemistry, other chemical theories were developed: Stereochemistry, Chemical Kinetic, Chemical-Physics, Spectroscopy, etc. How many chemical theories? Let us select the most representative ones by exploiting again the four pairs of basic choices. First, Classical Chemistry is characterized as PI and PO for the following reasons. Its organization is not the deductive one, AO, since it is not derived from a priori principles, but is based on a problem ("how many elements of matter are there?") and moreover most classical chemists - in particular, Mendeleev -, wrote texts in which they made use of DNPs, hence non-classical logic. In addition, Classical Chemistry makes use of a very elementary mathematics, at most rational numbers; hence, PI. Second, the theory of Physical Chemistry derives from a non-chemical, but physical, notion of entropy, which therefore plays the role of an axiom; as a consequence, the organization of the theory is AO; moreover, it makes use of the same elementary mathematics as Thermodynamics, which no physicist would consider on a par with the second order differential equations of Mechanics.[15] In sum, a PI&AO theory. Third, a minor, but very relevant, theory, i.e. Kinetic Chemistry. This theory is characterized by the choice AI, since it introduced the calculus into a field, Chemistry, which almost entirely ignored higher mathematics; certainly, its differential equations are only of the first order (and surely their solutions have constructive counter-parts), but at the time of its birth this theory was counted among the highly mathematized theories owing to its including infinitesimal analysis. Moreover, its organization is aimed at solving a problem, i.e. to studying the crucial notion of this theory, reaction speed. Hence, it is an AI&PO theory.

According to the characterizations - given in Sect. 6 - of the prime principles of reasoning the prime principle of Classical Chemistry is limitation, which in fact is constituted by the valence of each atom. The prime principle of Physical-Chemistry is that of attributing (chemical) existence to a mathematical (and even physical) notion, i.e. entropy. The prime principle of Kinetic chemistry is the extremant; indeed, this theory is aimed at accelerating as much as possible; the speed of chemical reaction. About a fourth prime principle, notice that at Peirce's time chemists applied the prime principle of causality by reasoning about classical Mechanics, to which most of them - e.g. Mendeleev (Mendeleev 1905, I, xi, n. 2) and the same Peirce referred with the aim of obtaining an ultimate explanation of the foundations of Chemistry (Peirce wanted to explain the foundations of Chemistry through the notions of form and mechanical ("Boscovichian" 7.509) force; 1.428).

The first consequence, at Peirce's time, was that there existed three chemical theories that suggested reasoning according to three different WoRs; moreover, a chemist had unavoidably to refer to one more WoR, that of the paradigmatic theory of Mechanics, which relied on the choices AI and AO (the same choices of subsequent Quantum chemistry) whose prime principle is causality, hence deduction.

[15] Van't Hoff equation together with Maxwell' relations among thermodynamic potentials - all first order differential equations – concern magnitudes of physical import, which are assumed as a priori axioms by a chemist.

The second consequence was that CT's use of all WoRs at the same time is not a historical novelty, because already a century before CT the professional practice of chemists introduced this mode of reasoning.[16]

9 The Origin of Peirce's Investigations on WoRs

In total, Chemistry presents four WoRs, in agreement with the four prime physical principles. As a chemist, Peirce was familiar with all the four WoRs that a chemist of his time took into account (although he applied the WoR of deduction in Mechanics, rather than in Quantum Mechanics)[17].

This fact explains: (*i*) His insistence on his going beyond the causality principle alone, i.e. the deductive WoR that the mainstream attributed to Mathematics in its entirety and to paradigmatic Physics, i.e. Mechanics. (*ii*) His great interest in discovering all possible WoRs. (*iii*) His insistence on suggesting, beyond the two commonly accepted inference processes, i.e. deduction and induction, one more inference process, abduction. (*iv*) At present time we know that this inference process is the prime principle of a very important theory, Physical Chemistry, in which he was the first Ph. D. at Harvard University; with respect to the empirical solutions given by classical chemistry the solutions obtained by Physical Chemistry represent the answers coming from an oracle, since they are essentially derived from the notion of entropy which is not of a chemical, but thermodynamic nature. Peirce revealed that his source of his ideas about the inference processes was Chemistry and in particular Physical Chemistry when he wrote:

> "By a hypothesis [read: abduction] I mean…., it is merely a supposition about an observed object… but also any other supposed truth from which would result such facts as have been observed, as when van 't Hoff, having remarked that the osmotic pressure of one per cent solutions of a number of chemical substances was inversely proportional to their atomic weights, thought that perhaps the same relation would be found to exist between the same properties of any other chemical substance."(6.254f.).

(*v*) His belief that it is possible to accurately distinguish among all the inference processes, although he lacked evidence to support his thesis.

[16] By pertaining to a different level of explanation – i.e. electrons - from the chemical ones, Quantum Chemistry introduced the same argumentative habit of a physical theory, i.e. reasoning according to one principle, in this case deduction, leading to regarding as primitive the other (chemical) principles. Indeed, Quantum chemistry is an AO theory since it depends on Quantum mechanics which, not being a chemical theory, works as a principle-axiom; its mathematics (recall Dirac's δ-function) is higher than PI mathematics; hence, its choice is AI.

[17] In my opinion Peirce alluded to this situation when he wrote "Modern methods [of reasoning] have created modern science and this century has done more to create new methods than any former equal period." (7.61) Moreover, Fann (1970, p. 23) adds that "he maintains that the producing of a method for the discovery of methods is one of the main problems of logic." (3.364) "Though other definitions of logic occur in his writings, the one he used in his teaching at John Hopkins University (1879–1884) was that it is the art of devising methods of research, "the method of methods"." (7.59).

Yet, Peirce was unsuccessful in accomplishing his general program aimed at accurately defining the WoRs. Unfortunately, he was misled by his program of research into the foundations of Chemistry, looking for a mechanical explanation based on the notion of force (although at his time Energetists tried to dethrone the notion of force in favour of the notion of energy). Moreover, given that at Peirce's time Chemistry was a young theory, no clear-cut borderline distinguished the different chemical theories; and, in addition, since in each chemical theory it was not clear what distinct roles are played by mathematics and principles, the above distinctions among the different chemical theories have to be taken as simply perceived, but not clearly understood.

In addition, Peirce's WoRs on chemical elements were not similar to Mendeleev's. As a matter of fact, Peirce meditated for a long time on the classification of chemical elements, so that he claimed to have suggested a table like Mendeleev's (7.509). Yet, Peirce's expressed even differing evaluations of Mendeleev's Table which he seemed to be "in considerable doubt", 7.222). He rarely refers to the WoRs employed for constructing this table; moreover, in such cases he mentions an undefined induction (once he writes "pure induction"); but never abduction![18]

10 Conclusions

In conclusion, the two dichotomies have suggested structural categories of the reasoning. Through them we have mutually compared (*i*) Peirce's inference processes, (*ii*) CT's ways of formal reasoning through mathematical and logical tools, which are of an operative, hence objective, nature, (*iii*) the physical prime physical principles which are of both an intuitive and a mathematical nature. We have characterized the correspondences of all of them. We have obtained a characterization of the structure of the WoRs and, through of CT's, a formal qualification of this fourfold structure of the principles of reasoning.

We remarked a basic difference between a physical theory which makes use of one prime principle, and CT, which makes use of all WoRs at the same time. On the other hand, an examination of chemistry shows that it also makes use of all the WoRs at the same time. The fact that Peirce was educated as a chemist and hence made use of all WoRs at the same time explains his philosophical effort to inquire into inference processes, in particular the notion of an abduction, i.e. the characteristic WoR of the theory of Physical-Chemistry, in which Peirce was very competent.

More than a century after his effort, one may suggest that his profound genius in discovering very deep ideas concerning WoRs was facilitated by his implicit reference to a scientific field largely unknown and disregarded by the mainstream, i.e. the various chemical theories. However, we saw that in general Peirce had great difficulties in achieving clearly defined results (apart from his invention of mathematical recursion); he recognized only three of his four WoRs and never accurately defined them (apart from deduction). Any way, all the correspondences (listed in Table 2) between his

[18] Rather, he wanted to found a new theory of thinking which associates ideas in a similar way to the chemical way through which chemical elements associate themselves in compounds. It was called Phanerochemistry; later it became Semiotics.

investigations and CT's characteristic features lead us to qualify Peirce as the father not only of the philosophy of CT but also of its multiple manner of reasoning.

Acknowledgement. I thank Prof. David Braithwaite who corrected my poor English.

References

Achinstein P (1993) How to defend a theory without testing it: Niels Bohr and the "logic of pursuit". Midwest Stuides Philos. 18:90–120

Bishop E (1967) Foundations of constructive mathematics. McGraw-Hill, New York

Carnot L (1803) Principes fondamentaux de l'équilibre et du mouvement. Deterville, Paris

Chellas BF (1980) Modal logic. Cambridge University Press, Cambridge

Davis M et al (1995) Computability, complexity, and languages: fundamentals of theoretical computer science. Academic Press, New York

Drago A (1988) A characterization of Newtonian paradigm. In: Scheurer PB, Debrock G (eds) Newton's scientific and philosophical legacy. Kluwer Academic Press, Boston, pp 239–252

Drago A (1989) La rivoluzione francese ha realizzato il "programma giacobino" di rifondazione di tutta la scienza. In: Società Italiana Progresso Scientifico: Atti della XL Riunione; L'età della rivoluzione e il progresso delle scienze, Bologna, pp. 335–342

Drago A (1994) The modem fulfilment of Leibniz' program for a Scientia generalis. In: Breger H (ed), VI International Kongress: Leibniz und Europa, Hannover, pp. 185-195

Drago A (1996) Alternative mathematics and alternative theoretical physics: the method for linking them together. Epistemologia 19:33–50

Drago A (2004) A new appraisal of old formulations of mechanics. Am J Phys 72(3):407–409

Drago A (2011) I quattro modelli della teoria meccanica. In: Toscano M, Giannini G, Giannetto E (eds) Intorno a Galileo: La storia della fisica e il punto di svolta Galíleiano, Guaraldi, Rimini, pp 181–190

Drago A (2012) Pluralism in logic: the square of opposition, Leibniz' principle of sufficient reason and Markov's principle. In: Béziau J-Y, Jacquette D (eds) Around and beyond the square of opposition. Birkhaueser, Basel, pp 175–189

Drago A (2013) The emergence of two options from Einstein's first paper on quanta. In: Pisano R, Capecchi D, Lukesova A (eds) Physics, astronomy and engineering. Critical problems in the history of science and society. Scientia Socialis Press, Siauliai, pp 227–234

Drago A (2014) A logical model of Peirce's abduction as suggested by various theories concerning unknown objects. In: Magnani L (ed) Model-based reasoning in science and technology: theoretical and cognitive issues. Springer, Heidelberg, pp 315–338

Drago A (2015) The four prime principles of theoretical physics and their roles in the history. Atti Fondazione Ronchi 70(6):657–668

Drago A (2016) Defining Peirce's reasoning processes against the background of the mathematical reasoning of computability theory. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology: logical, epistemological, and cognitive issues. Springer, Heidelberg, pp 375–398

Drago A (2017) Koyré's revolutionary role in the historiography of science. In: Pisano R, Agassi J, Drozdova D (eds) Hypotheses and perspectives in the history and philosophy of science: homage to Alexandre Koyré 1892–1964. Springer, Heidelberg, pp 123–141

Drago A, Manno SD (1989) Le ipotesi fondamentali della meccanica secondo Lazare Carnot. Epistemologia 12:305–330

Drago A, Pisano R (2000) Interpretazione e ricostruzione delle Réflexions di Sadi Carnot mediante la logica non classica. Giornale di Fisica 41:195–215 (English translation in Atti Fond. Ronchi, 59 (2004), 615–644)
Dugas R (1950) Histoire de la Mécanique. Dunod, Paris
Dummett M (1977) Elements of intuitionism. Clarendon Press, Oxford
Einstein A (1905a) Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. Annalen der Physik 322(6):132–148
Einstein A (1905b) Letter to Conrad Habicht, April 14th
Fann KT (1970) Peirce's Theory of Abduction. The Hague, Mouton
Grize J-B (1970) «Logique», in Logique et la connaissance scientifique. In: Piaget J (ed) Encyclopédie de la Pléyade. Gallimard, Paris, pp 135–288
Grzegorczyk A (1964) Philosophical plausible formal interpretation of intuitionist logic. Indagationes Mathematicae 26:596–601
Horn L (2002) The Logic of logical double negation. In: Kato Y (ed) Proceedings of the Sophia Symposium on Negation. University of Sophia, Tokyo, pp 79–112
Horn L (2010) Multiple negations in English and other languages. In: Horn L (ed) The expression of negation. de Gruyter, Mouton, pp 111–148
Kuhn TS (1977) Mathematics versus experimental tradition in the development of physical sciences. The essential tension: selected studies in scientific tradition and change. University of Chicago Press, Chicago, pp 31–65
Lavoisier AL (1862–1892), Oeuvres de Lavoisier. Imprimerie Imperiale, t. 1, Paris
Leibniz GW (1710) Preface theodicy. Routledge, London
Lobachevsky NI (1955) Geometrische Untersuchungen zur der Theorie der Parallelinien (orig. 1840). Berlin, Finkl. (English translation as an Appendix in Bonola R. (1955), Non-Euclidean Geometry. Dover, New York)
Markov AA (1962) On constructive mathematics. Trudy Mathematichieskie Institut Steklov 67:8–14 English translation: 1971, Am Math Soc Trans 98(2):1–9
Mendeleev DI (1889) Faraday lecture. In: Il sistema periodico degli elementi. Tecnos, Rome 2004, pp 99–127
Mendeleev DI (1905) Principles of Chemistry (orig. 1871). Longmans, New York
Mendeleev DI (2004) Il sistema periodico degli elementi. Teknos, Roma
Newell A, Simon H (1967) Computer science as an empirical inquiry: symbols and search. Commun. ACM 19(3):126–133
Peirce CS (1881) On the logic of numbers. Am J Math 4:85–95
Peirce CS (1887) Logical machines. Am J Psychol 1:165–170
Peirce CS (1968a) Questions concerning certain claimed human faculties. J Speculative Philos 2:103-114
Peirce CS (1868b) Some consequences of four incapacities. J Speculative Philos 2:140–157
Peirce CS (1869) Grounds of validity of the laws of logic: further consequences of four incapacities. J Speculative Philos 2:193–208 All collected in (Peirce 1951–1958, vol 2)
Peirce CS (1958) Collected paper of Charles S Peirce, vol 8. Harvard, Cambridge University Press, Cambridge, p 1
Prawitz D, Melmnaas P-E (1968) A survey of some connections between classical intuitionistic and minimal logic. In: Schmidt HA, Schütte K, Thiele H-J (eds) Contributions to mathematical logic. Elsevier, Amsterdam, pp 215–229
Reichenbach H (1938) Experience and prediction. Chicago University Press, Chicago
Scerri ER (2007) The periodic table: its story and its significance. Oxford University Press, Oxford
Troelstra A, van Dalen D (1988) Constructivism in mathematics, vol 1. Elsevier, Amsterdam



Remarks on the Possibility of Ethical Reasoning in an Artificial Intelligence System by Means of Abductive Models

Alger Sans(✉) and David Casacuberta

Philosophy Department, Universitat Autònoma de Barcelona, Barcelona, Spain
{alger.sans,david.casacuberta}@uab.cat

Abstract. Machine learning and other types of AI algorithms are now commonly used to make decisions about important personal situations. Institutions use such algorithms to help them figure out whether a person should get a job, receive a loan or even be granted parole, sometimes leaving the decision completely to an automatic process. Unfortunately, these algorithms can easily become biased and make unjust decisions.

To avoid such problems, researchers are working to include an ethical framework in automatic decision systems. A well-known example is MIT's Moral Machine, which is used to extract the basic ethical intuitions underlying extensive interviews with humans in order to apply them to the design of ethical autonomous vehicles.

In this chapter, we want to show the limitations of current statistical methods based on preferences, and defend the use of abductive reasoning as a systematic tool for assigning values to possibilities and generating sets of ethical regulations for autonomous systems.

Keywords: Machine learning · Ethics of AI · Abduction · Values

1 Introduction

This chapter aims at presenting a series of theoretical considerations in order to defend the plausibility of using abductive reasoning as a way to develop real ethical interactions with machines. To that end, in the second section we will argue that finding a way to infuse ethics in digital protocols is not just something we may need in the future, when the so-called singularity will start, but a pressing issue in the present. The third section is designed to clarify key concepts such as what is value and to distinguish human norms from computer rules.

In the fourth section, we describe the main methods used by AI researchers these days to introduce ethical reasoning in computer systems, and argue that these methods and techniques are clearly not sufficient to capture real values and develop inferences from them. We advise against the use of different paradigms, such as the MIT's Moral Machine, etc., due to its confounding of values and preferences, and we show how emotions, being key elements to understand ethical values, cannot be properly modeled within a supervised machine learning paradigm.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 318–333, 2019.
https://doi.org/10.1007/978-3-030-32722-4_19

In the last section, we outline the main theoretical features that make abduction a very relevant candidate for the development of future innovative procedures for designing ethically reasoning digital systems. To do so, we show how possibilities and values are strongly interconnected in abductive models and how abduction can help us to go beyond dichotomies such as formal logic versus psychology or fact versus value.

2 Ethical Issues in Current Applications of Machine Learning Technologies

A relevant deal of the literature on the ethical implications of AI is focused on the dangers of "the singularity" or "superintelligence". What will happen when artificial beings become more intelligent than humans and therefore able to build other machines that are also smarter than we are? Unfortunately, one does not have to wait for such superintelligences to rise in order to worry about the ethical issues surrounding the deployment of contemporary "weak" artificial intelligence. While we will not be discussing this issue in depth here, the interested reader may refer to [7, 8].

Other relevant research is devoted to more theoretical issues, such as being able to formalize specific ethical knowledge. SIROCCO [30] uses the Ethics Transcription Language to capture how ethical explanations in engineering are made, using the ethics code of the National Society of Professional Engineers as background. Jeremy [7] formalizes a specific representation of a utilitarian ethical model in which the program tries to calculate in a hedonistic way to maximize pleasure for a majority of the agents implied in the dilemma. More recently, GENETH [3] has been arguing how such algorithms have to be based on ethical principles determined through a consensus of ethical experts, and those principles can be used to find further regulations which could be arranged to create autonomous ethical systems.

Another research venue is to use logic programming as a way to analyse and generate ethical argumentations. In [48] we are presented with a logical and abductive model to develop ethical judgements in agents able of intention or recognition among other basic cognitive abilities.

We strongly agree with such perspectives and consider this the right strategy to produce one day artificial systems able to generate sound ethical reasoning. In [45] the interested reader will find a systematic description and presentation of this research paradigm. Unfortunately, other approaches based on statistical inferences to produce preferences are gaining popularity.

When trying to code an ethical framework for AI, the most common paradigm is to collect user's preferences to create ethical rules. A good example of this very popular approach is MIT's "Moral Machine" project, in which researchers use the paradigm brought about by Philippa Foot's "Trolley problem" mental experiment [16] in order to grasp the basic intuitions underlying the participants' application of their own scale of values. In these studies, human participants are asked to make their values and preferences explicit in front of a hypothetical dilemma. This is about an autonomous car that has to either swerve left and strike an 8-year old girl or swerve right and strike an 80-year old grandmother and, given the car's velocity, either victim will be killed on

impact. Based on those explicit preferences, researchers aim to find some basic ethical rules for autonomous cars to use in the future.

As we have shown above, this is not the only paradigm to create ethical models for AI, but it is so pervasive that we consider very important to discuss why it is a problematic approach. We find this approach incomplete for two main reasons. The first one is both theoretical and methodological: some researchers employing the "Moral Machine" paradigm seem to confound values with preferences. We will devote more space to this problem in Sect. 4 of this paper. To see other criticisms and defenses of the "Moral Machine" model see [6, 17, 20, 25, 27].

The other reason is practical: we do not have to wait for the autonomous car to be common in our cities to start worrying about moral problems in AI. Some problems are starting to come up already. Many of them are related to the building of profiles on users and their use in making automatic decisions about them, especially when these decisions may affect their fundamental rights. As more and more data are collected about us, companies and governments use that information in order to automate decisions. Whether it is selecting the better candidates for a job or a university, deciding if one should be granted a loan, or even determining if a suspect of a crime is eligible for parole or has to wait in jail for trial. Algorithms programmed with machine learning techniques make such decisions in an autonomous way, with very small human intervention, sometimes none at all.

The same algorithms also filter our access to information and can generate filter bubbles [39, 41] that make us think everybody else believes in the same things and has the same values we do, giving us a false understanding of our society and the world around us, as well as raising prejudices against people with different worldviews and making them appear irrelevant. This kind of profiling helps to spread fake news and makes them more effective, because they can be used to connect with our very own fears and prejudices, as happened with Cambridge Analytica, the company that created specific news items to convince people to vote for Brexit using profiles obtained from their Facebook activity [10, 46].

Why are such pieces of software problematic? On the one side, because they may repeat actual unjust decisions and amplify them. Machine learning algorithms capture regularities and patterns from the external world. If judges in real life are biased and tend to deny parole to black people, then algorithms will learn from such situation and repeat the pattern [53]. If these new decisions, automatically made by algorithms and already biased, then are used to feed the algorithms again. Thus, the bias will be indefinitely preserved and maybe even amplified [28]. Those biases cannot be removed by simply excluding the race variable because the algorithm will produce similarly discriminatory results by picking up the same patterns from the interaction of associated variables such as neighborhood and occupation.

Biases can also come for an unrepresentative database. An important number of face recognition algorithms are unable to recognize people of African origin because the databases used to train those algorithms present an overwhelming majority (almost 80%) of people with lighter skin [9].

On the other side, machine learning algorithms are good at finding relevant correlations, but cannot distinguish those based on real causal relationships and those that are spurious [44]. Therefore, such algorithms may find out that divorcees are usually

less able to pay back their loans and end up systematically rejecting their applications even if there is no real causal relationship at play. Even if it is statistically true that many divorcees do not pay back their loans, it would be unjust to reject an application just because of the applicant's civil status. This has been tested empirically in [5] in which subjects were presented with decisions made by algorithms in several scenarios and were able to identify unjust decisions, pointing at arbitrariness, generalization, and indignity as the source for such a perception.

Therefore, two different goals when automating decisions collide. On one side, we expect decisions about paroles or credits to be fair. On the other we also expect them to be correct. In order for human beings to establish the fairness of a decision, it has to be simple enough for humans to follow the reasoning. However, as stated in [26] if we tinker with our machine learning models to improve their accuracy, models will become more complex and therefore more difficult to interpret by humans.

3 Preferences are not Values

We will understand "value" in the philosophical classical point of view as a *property that makes one object or fact better than another from a no quantitative criterion*. By quantitative criterion, we mean the possibility to decide something based on an ordinally defined scale. Any defined scale can be composed by values, precisely by the fact of to be part of a defined scale. In other words, defined scale voids the totality of the value for the benefit of the generalization. This last case shows the *transformation from norm to rule*. Though "norm" and "rule" come from the same semantic field, we can draw some distinctions: first of all, it is possible to generate personal norms ("personal" is not the same that "private") but it is impossible in the case of rules because they are generalizations. Another difference is the difficult to postulate norms like rules. Norms contain psychological elements that are lost when transformed into rules. Among these psychological elements, we are interested in the kind named *ethical values*.

We understand "ethical value" in the sense that some human specific actions towards certain objects are better than other actions toward other objects. That is, they are more preferable because they are understood as the correct ones (by understanding "correction" in an absolute or relative sense we can really determine the weight of ethical values). For that matter, we will understand it in its relative sense that there are certain norms that we understand as ethical, which we try to preserve with our actions. Thus, these actions are always human.

On the other hand, it is accepted that any action is in relation to something. Aside from the complex problem about the meaning and essence of "relation", we understand this concept like *an action that is exerted reciprocally between two or more objects*. In short, *interaction*. The key of our definition is "reciprocally" which implies equal or mutual correspondence. We need to pay attention to this detail to avoid the mistake of stating that using a thing implies a type of interaction (this would eliminate the *mutuality*). Besides that, we consider that this definition of "relation" does not exclude other concepts as "communication", but implies to start one-step before it.

All this implies that there is a gap between norms and rules and, therefore, there is a bigger gap between human norms and computational rules (even if these rules contain

human norms as human regulations). The case of computation and ethical values is the perfect example of *transformation from norm to rule* above cited. First, in a machine, any possibility is previously programmed. Secondly, this shows us that although all regulations are a type of rule, not every rule is a regulation. Here, regulation is the attempt to order the ethical aspects into the formal commands that compound an interactive machine. However, ethics can contemplate no regulation and still be in the system. For example, human relations are not laws but are the possibility of laws, and not the other way around. In terms of this comparison, if the will is to achieve a real moral AI -that is, a system that can be used in real life ethical problems, instead of one just designed to use a set of abstract principles to solve an intellectual puzzle-, then it is necessary to work around the idea that there are some things outside the system, as well as the code of the machine. A first obvious element outside the algorithm parameters is the human that interacts with the algorithm. This is the same case as the comparison between ethics and laws, because it is necessary to obtain such real interaction without losing ethical aspects. "System" has to be understood as something more than *some internal rules, principles or things*. Instead, we have to accept the idea of a more comprehensive sense that implies the external elements, which often are the possibilities of the internal sense of the system.

So, from a pragmatist approach, we can state that human actions and values need to be included in the system, and from here it follows that if our goal is to obtain real interaction, then it is necessary some kind of computational reasoning/action based on ethical principles.

Let's consider first a very basic digital interaction, like the ones coming from e-commerce. Interaction is mostly automatic and grounded on the idea of fair play of the participants. However, when some injustice is generated in the system, such as a provider maliciously sending a damaged product, or a hacker stealing the identity of a real customer, using the credit card of the legit user to buy stuff for, then the problem has to be handled by human operators. Some e-commerce platforms are designed in a way that such a process is easy, but most of the current service providers tend to make the process difficult hoping most people will abandon and just accept the unjust behavior.

We can divide this problem into two aspects. (1) On the one hand, when a person interacts with a system, it is simply a logical space in which nothing can be added, as I just said, the value is something human that is added to the things of the world [55, p. 183, 6.41]. If there is a serious problem, in the end, it is another person who has to solve it. But getting a person to solve this problem involves an arduous and tiring task. Therefore, ethical problems are always seen as an anomaly external to the system. When solved, only a specific case is solved and this can be repeated *ad infinitum*. In addition, the system becomes a race of obstacles, in which only the most persistent people reach the solution.

(2) On the other hand, it is found that there is *no real interaction* between the system and the person since the solution is given by another person, without affecting the machine. This situation is economically harmful and time-consuming, and leads to a false interaction in the sense that a person has used the system, but there has not been -or he has not *felt* - reciprocity.

When one uses a system and all works fine, one does not think about the interaction process. Only when problems appear one looks for real interaction [14]. This situation causes different emotions, none of them good. One of the most important of them is to "understand" that one is not interacting with anyone, simply using a system. In the end, it is a person who solves the problem in a very specific way. However, what is important here is that (1) there cannot be a rule that solves this because the system is not designed to take into account the actual interaction with people. In other words, it cannot adapt to contingent demands. On the other hand, regulations have to be created that (a) preserve the rights of the people beforehand and (b) draw an alternative path to the system. Therefore, most of the regulations serve to preserve ethical values (ethos) that can be lost during the interaction with the machine. However, there is a gap between the regulations aimed at preserving the ethical values and the ethical values that are to be preserved. In other words, this kind of rules simply determine what is to be protected, but are only achieved through interaction with another person (the final solution). Better said, the values are recovered when, in the end, a person solves the incidence that affected the customer.

The problem is not ethical, but technical. Amorality lies in the fact that the solutions are always external to the system, and the protocol of the machine and the sequences and orders of the employees. As we stated at the beginning of this section, ethics is about human action through a moral code. Thus, immorality is created because there is no real interaction between the customer and the system with which it is dealing. This generates three problems:

- Ethos cannot be preserved.
- The system does not participate in anything.
- The regulations only serve so that there is a limited number of incidents and, therefore, of amendments.

These regulations are the rules designed to keep the ethos of the project. However, there is a gap between the ethos and the regulations. The main challenge begins when it is understood that regulations cannot modify the rules of the system because there is no real interaction and, finally, this is a precondition for the possibility of ethics. The main challenge begins when it is understood that regulations cannot modify the rules of the system. The starting point should be the rules that present patterns that we identify as moral. In addition, this is not an isolated study of emotions and machines, but of humans and their feelings.

This begs the question as when somebody is ethic. Of course, in their actions when interacting with the world. But there is another element that must be added because it is not true that there is one ethical description for every ethical action. Otherwise, there would be no real ethical problems in general. On the contrary, it is generally assumed that no ethical descriptions are possible nor from natural objects (naturalistic fallacy) nor from idealist objects (impossible to be real in existence-natural terms). However, it is a human context phenomenon, that is, social. Likewise, society is not a finalized idea, ethics are unfinished because through ethical actions we construct the society that we want. From this idea, recognize some action as ethical is a social context familiarization question and, finally, that bears with human feelings.

A way to understand this is using the concept of *form of life* by the second Wittgenstein [55]. According to Wittgenstein agreements on how to use language within a group are only possible if everybody in such a group shares a common form of life. We all understand the word "red" in the same way because we are all humans and we share the same physiological system to recognize colors. Paraphrasing Wittgenstein, if a bee could talk about colors, we wouldn't understand it, because its system to perceive colors is very different from ours. Following such a train of thought, we could say that it is difficult to take seriously someone affirming that some ape is a painter by the fact that it can paint something. There are some arguments against this idea but, in the end, all of them lie in one elemental thing: apes are not humans, and painting is a human activity[1]. In the same way, it is difficult to consider an AI's painter[2] as an artist now[3].

In the same vein, the value problem is not about the perception of it, but of recognizing the entity that bears it as moral. To establish this, it is necessary something more than the ability to do actions labeled as moral, it is necessary (a) to arrive to such actions through moral reasoning, and (b) to realize the individuality of such solutions: the ways of solving a problem that may seem equivalent for an external observer will probably generate different appreciations depending on the subject that has to develop them.

This problem comes in two dimensions, one is material as we do not have the capacity to really emulate the way human process decisions, and the other is formal as we do not have a way to convert ethical reasonings into formal generalizations. A possible solution may be to apply some sort of abductive reasoning to reduce such gaps, a possibility we will explore in the following sections.

4 Beyond the Moral Machine

Let us consider now the "Moral Machine" from the point of view of what we have discussed in Sect. 2. We find that it is clearly a case of a "conversion process" in which personal norms are transformed into general regulations [6, 19]. When users of the system are invited to decide which type of person is more important and therefore should be preserved, they rely on different intuitions based on their experience as human beings -on their forms of life, Wittgenstein would say. This covers their general preferences about who they like and dislike, what is considered more valuable in current times, as well as emotions, religious beliefs and so on. If in a Socratic style, we enquired these users about why they made such decisions, and which ones were ethical, we would find that not all their preferences are ethical ones. To see a more nuanced philosophical development of the philosophical stance of separating values and preferences see the examples [18] and [11].

[1] For example, there is the authors' rights problem in cases like Naruto's (macaque) selfie (monkey selfie copyright dispute).

[2] An example is GANS' Portrait of Edmond Belamy sold in 2018 for $432,500.

[3] Both the first and the second examples can be compared and evaluate through some random pseudo-infinite monkey theorem.

The authors of the original research on "Moral Machine" have pointed that out in different places. For example, in an interview for the Medium Platform, they described a very immoral tendency among certain users to "go with the flow" of our neoliberal society and decide that it is always better to kill a homeless person than an executive [37]. They also concede the point made by Dreyfus [15] that there is such a thing as moral expertise, and therefore the opinion of ethical experts should probably be considered more relevant than that of the general public when deciding which rules to implement, the same way that a company will trust the expert judgment of an engineer in order to see if a technology is safe.

Therefore, we can conclude that The Moral Machine approach cannot infuse autonomous cars with real ethical judgment but just with a messy bunch of preprogrammed intuitions in which it will be very difficult to distinguish ethical values from emotional preferences. In addition to that, what we get is just a scale of values unable to capture the psychological and philosophical nuances of our ethical preferences. We must conclude, then, that such a quantitative approach will not be able to solve the problem of the biased behavior of machine learning algorithms we described in Sect. 1, nor the simpler problem of putting real ethical reasoning into an e- commerce interface presented in Sect. 2.

As we said before, machine learning is the basic paradigm under which most artificial intelligence software is developed today. Some algorithms use supervised methods in which the proper result is "taught" to the algorithm by giving it examples that help to link specific inputs to a specific output. In unsupervised learning, the algorithm receives raw data and tries to organize them in clusters based on the similarity of the inputs. In reinforcement learning, the software improves its outcomes by comparing them to a desired final result, and it is fed "rewards" and "punishments" in order to change the way inputs are processed. Most of the commercial software used to make the kind of automatic decisions we discussed in Sect. 1 (job interviews, access to university, loan applications, along with others) is prepared with the supervised learning methodology. When developing COMPAS, and other pieces of software used to decide whether a person should be granted parole, developers use a list of the characteristics of those previously processed by the justice system, including their gender, age and address, as well as former offenses, the type of the crime they committed, and so on [53]. This input is then associated with an existing answer, the output- whether those people were granted parole or not. The algorithm looks for patterns that connect the input to output and tries both to predict new inputs -whether that person will commit a crime under parole- and to decide the output - whether it is a wise idea to send them back to the streets.

Suppose that we want to add emotions to the deliberative process of the algorithm. If we follow Wittgenstein's ideas about forms of life, it should be clear that exposing the system to just some elements of the context that surrounds the decision will not be enough. We need a very close understanding of what it means to be human to be able to convert the human form of life into something that can be processed properly by an algorithm.

The main way in which emotions are taught to computers and robots right now is using facial expressions: humans rate pictures of faces expressing different emotions - sadness, happiness, surprise, and so on- and then researchers use a supervised learning algorithm to find patterns that connect such data to specific emotional states. Facial

expressions data can be amplified by adding physiological information such as pulse rates, breathing rates, and skin conductivity.

Unfortunately, such an approach cannot give us a proper understanding of emotions, mainly because of two reasons:

(1) Humans are very bad at labeling emotions based only on facial expressions. We recognize emotions easily when portrayed by a good actor -that is, when they correspond to a code we have previously learned by watching movies and series, but are much less proficient at interpreting real life emotional expressions. The system we normally use to attribute emotions to an agent is far more complex [1]. When one is invited to look at a picture of a person expressing a specific emotion, it is common to misinterpret it, seeing sadness or anger when there was only surprise, for example. Judgments about the same facial expression tend to change when it is accompanied by a brief context, and a neutral face is interpreted as happy, angry or sad depending on the story that is presented with the picture [23]. Also, our own emotional states affect the way we interpret other people's emotions, and we see sadness in a neutral face if we happen to be sad that day [13].
(2) There is a big gap between the neurophysiological basis of emotion and the way our culture labels such neurophysiological data. So far, the only "objective" elements one can connect to our emotional states are valence and intensity [12, 55]. Some emotions are received in a positive way, others in a negative one, and one can find emotions from almost undetectable moods to explosions of deeply moving feelings. When we label some emotional state as disgust, rage or abhorrence our culture -our form of life- has a great influence. As the model of the constructed emotions first developed by Lisa Feldman Barret showed [4], the way our parents label our emotional states as babies and small children are the basis by which we build our own understanding of emotions. Whether a mental state is labeled as "guilt" or "shame" depends on how the words were introduced by our parents. Whether peas have a disgusting taste or they are frustrating just because we prefer French fries instead depends on the way our parents interpreted and then labeled our first interactions with that particular food.

If we take that into consideration, we can see how emotions, being a crucial part of our understanding of the ways humans make ethical decisions, cannot be adequately modeled by present AI models. Maybe when the singularity will come -if it comes at all- computers will be able to understand emotions even better than us, but then biases in decision-making processes will not be in the list of problems to be addressed.

5 The Abductive Way Towards Ethics

One tentative solution could be to amplify machine learning algorithms with abductive reasoning. An example of this idea is Magnani's mimesis concept [31], which is focused on understanding contemporary technological artifacts as devices to modify both AIs and humans in their interactive processes. However, the operating range of Magnani's proposal is provided by perception and representation considered in terms of some imitation game. To solve our problem, even if it is theoretical, it will be

necessary to provide some further arguments. This approach does not try to do some speculation about the possibility to program moral aspects but suggest a different perspective to address the present issue.

One possibility would be to research on understanding human internal ethical intentionality through the external ones, represented by models as moral mediators [34, 35]. So that the system would be able to abduce the psychology and the feelings of the users implied in the process, go beyond the patterns and correlations and unearth real ethical judgments instead. Considering this, technological artifacts and machines can be understood as Passive moral [35, p. 70] triggers on passive intellects, where is known that "passive" means "potency" (*intellectus possibilis*). Then, these technological objects would be able to distribute human moral capacity [35, p. 68].

Moral mediators are constructed above epistemic mediators [36, Ch. III], external representations and objects which are relevant for reasoning and discovery processes [35, p. 73]. The problem at hand is that there are tacit templates in human moral action and implicit patterns of behavior [34, p. 100], often understood like psychological aspects. Therefore, research should be on abductive process to recognize internal intentionality through these patterns of behavior, understood as external intentionality. The focus would be a good moral distribution.

However, this path is difficult. It is known that through abduction it is possible to show that there is no Frege's logic/psychology dichotomy [2, p. 65, 21, p. 17, 33, p. 287, 40, p. 378, 52, p. 7 et al.]. Parting from these observations and from the abduction concept itself it is easy to identify and collapse other dichotomies, such as justification/discovery (it is impossible to cite some relevant works about it because all contemporary debates around abduction start with this dichotomy) and fact/value [47]. From these three dichotomies, we can infer that our reasoning apparatus is both descriptive and value based. Therefore, our human capacities cannot be divided in logic and psychology, because the first is one possible representation of the second, which consists of descriptive and evaluative capacities. In conclusion, "we are both producers and consumers of knowledge" [21, p. 11].

This last sentence is closely linked to the pragmatist principles, and shows the way to solve the problem of the neglected view [2, p. 39, et al.] to which Hanson and Feyerabend were some of the most important contributors. Inspired by Hanson's theory-ladenness, Simon [51] wanted to refute Popper's *Logic der Forschung*[4] via computational science through showing the possibility to heuristic information description. It is interesting to note that both Hanson's theory and Feyerabend's epistemological anarchism use abduction (or counterinduction) to stress that reasoning is not only descriptive by necessity and probability. By contrast, the capacity to add possibilities is what is needed [49].

However, this is a very difficult task, and we would like to argue that it is not feasible within our current level of technological development. The results of computational attempts to characterize abduction are either (a) merely theoretical proposals

[4] The reason to put the title in German and not in English is that "Forschung" is "research", not "discovery". Thus, as Aliseda says [2, p. 12], English translation «The Logic of Scientific Discovery» is not correct.

or (b) not about abduction at all. An example of this last case can be found in Kakas' definition of abduction as deduction in reverse [24]. On the other hand, an example of the first is Thagard's IP [52], where it is assumed that there is a necessary link in terms of abduction between deduction and induction, and that this kind of reasoning is something more than logic (in a Fregean sense), something rather psychological [52, p. 56]. Again, the problem is that this psychological element is not identified. Without a compromise with any metaphysical or theoretical historical concept (which often are the same), it is possible to name it creativity in Łukasiewicz's sense [29] as there is no explanation without the creative capacity to generalize [49].

First of all, it is necessary to consider abduction in terms of value in perception [50] [5]or, in other words, to admit that our perception implies questions about values. For example, we decide for one theory and against another because of their perceived beauty. This reflection done by Putnam [47] illustrates how bridges can be built between the description of evaluations and the use of evaluations for descriptive purposes.

The existence of descriptive values is important because it allows us to show the role of values in the description of the world. But, while this role is important, it is not the only one. In short, only facts are descriptive. This means that Putnam's attack is directed against a false conception of neutrality based on a dichotomous logic. However, to say that values may be implied in the description of the facts that compose the world is not the same thing than to say that values play a necessary part in establishing which facts are possible in terms of which world we want. It is necessary to stress that when we want this world, we will focus our efforts on getting it. In other words, "as a form of 'possible worlds' anticipation" [34, p. 99].

The importance of this proposal lies in the fact that it shows how the imagining of possibilities implies the use of something else than factual descriptions. There may be a shifting balance: usually, factual descriptions may be more important than factual evaluations, but there may also be situations in which evaluations are more important than factual descriptions, and ones in which there is no room at all for factual evaluations. For us, this assumption is important because there are values in abductive reasoning in the presumptive sense, that is, there are propositions that express how facts should be analyzed. This is clear because abduction generates different possibilities for reasoning but, at the same time, these possibilities are the only ones, and so, in the end, one of them will be more accurate than others.

A possibility is not an object and, although it exists, its existence is not the same as the existence of objects. On the other hand, abduction is not an object either. It has something in common with possibility, but it also shares some features with rules. Possibility is, using a redundancy, possible by abduction. Abduction is the rule that enables the existence of possibilities. Furthermore, given that innovation is the materialization of a possibility (with other items) it is easy to conclude that abduction is the basis of innovation (without these items, of course).

This outline forces us to formulate the question about the method to follow when introducing a new rule in our ontological schema (in the sense of the entire set of data

[5] The reference is only for the "abduction as perception" argument.

that configures our worldview). The consideration of this matter is important because if our ontology consists of language-games related to worldviews, then abduction "exists" in the particular sense of something that had already existed before, but was hidden until now. But this explanation is not satisfactory, because what makes the "fundamental problem" [21, 22] contemporary [38] is precisely the addition of abduction to the debate. Our language is at the same time the possibility and the limit of our world [55, p. 149, 5.6] and this point is important because its determinacy or cluelessness configures our worldview by way of its relation to what we can do or not. However, this dependency can be understood in a pragmatic point of view.

From this point of view, the idea is simple. The tools that one uses build the world in which one lives, but this world cannot be the whole of reality: if no ontological relation with some general schema is needed, then we can suppose that this general schema is not necessary. Thus, we must conclude that abduction has not been "discovered", but it is nevertheless used and, as a consequence, our world is not the same as before -a new rule has been added to it. What happens when this rule is incorporated in a semantic world is shown by Magnani's EC-Model which is based on a pragmatic sense of "contextual" [32, 42, 43] very close to Wittgenstein's notion of language-games. This idea is reinforced by the fact that while abduction occurs every day, it is seen as an anomaly or as something impossible to explain.

In this sense, abduction as reasoning about possibilities can exceed descriptive capacity. This happens, on the one hand, because a possibility is not a given fact and, on the other hand, because possibilities are often only an expression of the world we want. This last point is important to our project because in moral contexts (that is, contexts in which ethical evaluation is more important than descriptions) propositions allow us to express the facts as we think they will turn out to be. Thus, we have to reconceive the meaning of abduction. From the point of view that innovation can be a material object, for instance, a performance exists inasmuch as it is an event. Moreover, an abduction schema must consider the idea that the cutdown[6] problem cannot be solved by comparing descriptions, and this is because abduction is about possibilities, and they do not necessarily have to be descriptions or knowable objects.

As a result, we want to advance the idea that abduction starts in an evaluative mode that includes everything we consider before deciding that something is possible. The action where we consider some possibility is an abduction. This entails the idea that abduction incorporates a new semantics in which we can present valuative descriptions that can interrelate with habitual descriptions without losing its qualities. But for this to be possible, we need descriptive evaluations of possibility, because we need to have a knowledge of possibilities that does not depend necessarily on objectification. By adding this semantic approach to abduction to the EC- Model, contexts can be understood as language-games without affecting their cognitive theory origins, and this is advantageous for us because it allows us to use evaluative semantics as a new tool to AI debates.

[6] Cutdown problem refers to two questions around abductive reasoning, the first is about the specification of the conditions for thinking up possibilities for selection, and the second is about the characterization of these possibilities. See also [32, pp. 6–10] and [42, p. 234].

6 Conclusions

Following the current discussion regarding values and abduction, we have argued how values cannot be comprised in a specific system of regulations. A system able to do that would need to get a fine and detailed grasp of what is being human, and opens the question about what are the specificities of the human form of life.

Any regulation built on the idea of a quantitative scale will fail to capture the fine-tune of human reasoning about ethics, which includes, among other stuff, emotions and feelings. Without such an understanding, engineering approaches such as The Moral Machine will not be able to model human ethical reasoning.

Current AI technologies are not yet able to capture all the nuances of human ethical reasoning. Systems based on the machine learning paradigm cannot escape the naturalistic fallacy, as they are trained by current examples of ethical decisions, that is, on what there is. As the biased behavior of programs such as COMPAS shows, such techniques may lead to unjust decisions in critical situations about our freedoms. Purely engineering approaches that try to create working models for AIs based on the rational choice theory from economics or explicit preferences expressed by humans such as the Moral Machine also fall in the same naturalistic fallacy, not realizing that there is an immense difference between values and preferences. Then, of course, there are AI models that just want to show how some formal procedures of ethical reasoning can be modeled and studied with computer simulations. We do not have any objections to such research and think they will help in the future when the task of producing computers able to understand ethical arguments and evaluate them will be properly understood.

Our proposal based on abductive reasoning is a theoretical answer to the lacks that we have shown. Abductive reasoning permits to consider a different human being dimension of reasoning, which is not usual yet. Epistemological and cognitive aspects of abduction have been drawn in cognitive and AI research, but not in all its aspects. There are several historical reasons which are not relevant here but, it is important to note that this reality implies an incompleteness of all this research. One of these implications has been described in this paper, that is, the fact that there is no real ethical research in AI.

To consider abduction as an essential part of ethical research is not nonsense at all. In fact, the pragmatist approach of this concept tries to acquire a solution to the synthetic problem about our participation in the process to apprehend (in Kant's meaning of Apprehension). In other words, the first step of *Weltanschauung* conformation. From a pragmatist point of view, in all of these processes, we add the possibility to combine different elements in novel ways to obtain our world image. The relevant contribution of pragmatism is to consider the idea that there are evaluative elements in this first synthetic apprehension, also known as perception. Then, from this theory, knowledge acquisition implies values. It is not difficult to find examples where these values imply also ethical aspects. This is shown by West, who offers another aspect of Peirce's third kind of reasoning:[7]

[7] The passage quoted by West is from Peirce.

> The third kind of reasoning tries what illume naturale, which lit the footsteps of Galileo, can do. It is really an appeal to instinct. Thus reason, for all the frills it customarily wears, in vital crises, comes down upon its marrow-bones to beg the succour of instinct [54, p. 476].

If it is possible to understand abduction in this sense, then it is possible to conceive that general knowledge implies valuations too. Furthermore, it is possible to speak about ethical knowledge without naturalistic fallacy commitment, because description and prescription are combined trough action -to know/perform the world-.

Therefore, we strongly believe that any attempt to produce AI models of ethical reasoning cannot be based only in engineering criteria and efficiency protocols, but need to include a philosophical stance, realizing that ethical arguments are a completely different type of arguments from those of science and logic, and require a completely different procedure to generate conclusions and produce insights. If we want to develop reasoning mechanisms to detect biases and injustices in automatic decision systems, abduction is a sound and promising model to do as it offers innovative mechanisms to link possibilities and values and go beyond classical dichotomies.

Acknowledgments. We would like to express our gratitude to TecnoCog research group: Anna Estany, Jordi Vallverdú, Dafne Muntanyola, and Rosa Herrera. On the other hand, it is necessary to express gratitude to Lorenzo Magnani and Atocha Aliseda because the part of abduction would have been impossible without their advice.

This research paper has been possible by the Research group Epistemic Innovation: the case of the biomedical sciences (FFI2017-85711-P) and FPU predoctoral program.

References

1. Adolphs R (2002) Recognizing emotion from facial expressions: psychological and neurological mechanisms. Behav Cogn Neurosci Rev 1(1):21–62
2. Aliseda A (2006) Abductive reasoning: logical investigations into discovery and explanation. Springer, The Netherlands
3. Anderson M, Anderson SL (2015) Toward ensuring ethical behavior from autonomous systems: a case-supported principle based paradigm. In: Proceeding of the AAAI workshop on artificial intelligence and ethics (1st international workshop on AI and ethics)
4. Barrett LF (2006) Solving the emotion paradox: categorization and the experience of emotion. Pers Soc Psychol Rev 10(1):20–46
5. Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) It's reducing a human being to a percentage: perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI conference on human factors in computing systems. ACM, p 377
6. Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293):1573–1576
7. Bostrom N (2014) Superintelligence: paths, dangers, strategies. OUP, Oxford
8. Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: The cambridge handbook of artificial intelligence, pp 316–334
9. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. Proc Mach Learn Res 81:1–15
10. Cadwalladr C, Graham-Harrison E (2018) The Cambridge analytica files. The Guardian 21:6–7

11. Callicott JB (2018) Ecological sustainability. In: A sustainable philosophy—the work of bryan norton. Springer, Cham, pp 27–47
12. Cunningham WA, Raye CL, Johnson MK (2004) Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. J Cogn Neurosci 16(10):1717–1729
13. Douilliez C, Yzerbyt V, Gilboa-Schechtman E, Philippot P (2012) Social anxiety biases the evaluation of facial displays: evidence from single face and multi-facial stimuli. Cogn Emot 26(6):1107–1115
14. Dreyfus HL (1997) Heidegger on gaining a free relation to technology. In: Technology and values, pp 41–54
15. Dreyfus HL, Dreyfus SE (2004) The ethical implications of the five- stage-skill-acquisition model. Bull Sci Technol Soc 24:251–264
16. Foot P (1967) The problem of abortion and the doctrine of double effect. Oxford Rev 5:5–15
17. Goodall N (2016) Away from trolley problems and toward risk management. Appl Artif Intell 30(8):810–821
18. Green SJ (1989) Competitive equality of opportunity: a defense. Ethics 100(1):5–32
19. Hevelke A, Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci Eng Ethics 21(3):619–630
20. Himmelreich J (2018) Never mind the trolley: the ethics of autonomous vehicles in mundane situations. Ethical Theory Moral Pract 21:669–684
21. Hintikka J (2007) Socratic epistemology. In: Explorations of knowledge-seeking by questioning. Cambridge University Press, Cambridge
22. Hintikka J (1998) What is abduction? the fundamental problem of contemporary epistemology. Trans Charles Sanders Peirce Soc 34:503–533
23. Jain AK, Li SZ (2011) Handbook of face recognition. Springer, Heidelberg
24. Kakas AC (2017) Abduction. In: Sammut C, Webb, GI (eds) Encyclopedia of machine learning and data mining. Springer, New York
25. Keeling, G (2019) Why trolley problems matter for the ethics of automated vehicles. Sci Eng Ethics 1–15
26. Kleinberg J, Mullainathan S (2018) Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. arXiv preprint arXiv:1809.04578
27. Lin P (2016) Why ethics matters for autonomous cars. In: Maurer IM, Gerdes J, Lenz B, Winner H (Eds) Autonomous driving: technical, legal and social aspects. Springer, Heidelberg, pp 69–85
28. Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. arXiv preprint arXiv:1803.04383
29. Łukasiewicz J (1970) Creative elements in science. selected works. North-Holland Publishing Company, Amsterdam, pp 1–15
30. McLaren BM (2003) Extensionally defining principles and cases in ethics: an AI model. Artif. Intell. J. 150:145–181
31. Magnani L (2018) Eco-cognitive computationalism: from mimetic minds to morphology-based enchancement of mimetic bodies. Etropy 20:430–446
32. Magnani L (2017) The abductive structure of scientific creativity. In: An essay on the ecology of cognition. Springer, Switzerland
33. Magnani L (2009) Abductive cognition. In: The epistemological and eco-cognitive dimensions of hypothetical reasoning. Springer, Heidelberg
34. Magnani L, Bardone E (2007) Distributed morality: exteralizing ethical knowledge in [β] technological 39 artifacts. Found Sci 13:99–108
35. Magnani L (2006) La moralidad distribuida y la tecnología. Cómo las cosas nos hacen morales (trans: Olmos P (UNED) and Feltrero R (IFS, CSIC)). Isegoría 34:63–78

36. Magnani L (2001) Abduction, reason, and science. In: Processes of discovery and explanation. Kluwer Academic/plenum Publishers, New York
37. Meaker M (2019) How should self-driving cars choose who not to kill? medium platform
38. Nepomuceno-Fernández A, Soler-Toscano F, Velázquez-Quesada FR (2014) The fundamental problem of contemporary epistemology. Teorema 23(2):89–103
39. Nguyen TT, Hui PM, Harper FM, Terveen L, Konstan JA (2014) Exploring the filter bubble: the effect of using recommender systems on content diversity. In: Proceedings of the 23rd international conference on world wide web. ACM, pp 677–686
40. Niiniluoto I (2014) Representation and truthlikeness. Found Sci 19(4):375–379
41. Pariser E (2011) The filter bubble: what the Internet is hiding from you. Penguin, London
42. Park W (2017) Abduction in context. In: The conjectural dynamics of scientific reasoning. Springer, Switzerland
43. Park W (2015) On classifying abduction. J Appl Logic 13(3):215–238
44. Pearl J, Mackenzie D (2018) The book of why: the new science of cause and effect. Basic Books
45. Pereira LM, Saptawijaya A (2016) Programming machine ethics, vol. 26. Springer, Cham
46. Persily N (2017) The 2016 US election: can democracy survive the internet? J Democracy 28(2):63–76
47. Putnam H (2002) The collapse of fact/value dichotomy and other essays. Harvard University Press, Cambridge
48. Saptawijaya A, Pereira LM (2015) The potential of logic programming as a computational tool to model morality. In: A construction manual for robots' ethical systems. Springer, Cham, pp 169–210
49. Sans A (2017) El lado epistemológico de las abducciones: La creatividad en las verdades-proyectadas. Revista iberoamericana de argumentación 15:77–91
50. Shanahan M (2005) Perception as abduction: turning sensor data into meaningful representation. Cogn Sci 29:103–134
51. Simon H (1977) Does scientific discovery have a logic? In: Models of discovery, Pallas Paperback, Holland, pp 326–337
52. Thagard P (1988) Computational philosophy of science. MIT Press, Massachusetts
53. Washington AL (2019) How to argue with an algorithm: lessons from the COMPAS propublica debate. Colorado Technol Law J 17(1)
54. West C (1989) The American evasion of philosophy. In: A genealogy of pragmatism. The University of Wisconsin Press, Madison
55. Wittgenstein L (2009) Philosophical investigations. Wiley, Hoboken

Epistemological and Technological Issues

On the Follies of Intercourse Between Models and Fiction: A Naturalized Causal-Response Diagnosis

John Woods(✉)

Department of Philosophy, University of British Columbia, 1866 Main Mall, Vancouver, BC V6T 1Z1, Canada
john.woods@ubc.ca
http://www.johnwoods.ca

> "I am clear that [the philosophy of mathematics] would be a fictionalist account, legitimizing the use of mathematics and all its intratheoretic distinctions in the cause of that use, unaffected by disbelief in the entities mathematics purports to be about."
>
> Bas van Fraassen (1980, p. 35)

> "Not all properties of convenience will be real ones. There are the obvious idealizations of physics – infinite potentials, zero time correlations, perfectly rigid rods, and frictionless planes. But it would be a mistake to think entirely in terms of idealizations – of properties which we consider as limiting cases to which we can approach closer and closer in reality. For some properties are not even approached in reality. They are pure fictions."
>
> Nancy Cartwright (1983, p. 153).

> "No theory of mathematical existence, let alone any resolution of any realism vs. anti-realism debate, will have any bearing on mathematics and its practice."
>
> Akihiro Kanamori (2013, p. 24).

> "The existence of fiction is a powerful argument for absolutely nothing ….".
>
> Saul Kripke (1973, p. 23).

Abstract. Efforts to model the unrealities of model-based science on the unrealities of literary fiction are unavailing. There are three main reasons for this. One is that the dominant theories of fictional stories are drawn from the philosophy of language which, in turn, tends to model reference, truth and inference in the way Tarski did in the model theory of first-order languages. In so doing, philosophy of language imports into theories of natural language the Basic Laws of Semantics, which provide that reference and ascription to the unreal is impossible. Reference to and quantification over the unreal is semantically empty and alethically barren. A second difficulty flows directly from the first. If unrealist science did succeed in modelling itself on the dominant theory of the literary unreal, the language of science would be faced with semantic and alethic collapse. A third difficulty is that literary discourse is possessed of a semantic and epistemic singularity that

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 337–371, 2019.
https://doi.org/10.1007/978-3-030-32722-4_20

has no imaginable counterpart in the language of science. The semantic singularity is that the sentences of stories are unambiguously and concurrently true and not true. The epistemic singularity is that every reader of stories knows this and isn't the least troubled by it. The reason why is that, although inconsistent, the sentences of stories aren't contradictions. There is nothing remotely like this in semantico-epistemic structure of model-based science.

Keywords: Ambiguity · Aristotle · Basic laws of semantics · Basic laws of fiction · Cartwright sentences · Contradictions · Creative realism · Fiction · Fictionalism · Formal semantics · Inconsistency · Meinongianism · Model theory · N-fic · Pretendism · Quantification · Reference · Reification · Truth-sites · Unrealism (anti-realism)

1 Provocations of the Unreal

Unrealist scientific theories are said to be theories whose domains contain unreal entities and whose working vocabularies contain expressions which denote them, further expressions which ascribe properties to them, and also the wherewithal to put those ascriptions to load-bearing inferential uses.[1] Models are a prominent feature of the sciences, sometimes used as investigatory aids, and are helpful in the creation of theories in which they themselves have no official place. Models exclusively of this kind have practical value. They help in getting a theory thought-up, but they are left on the post-creative cutting-room floor and make no subsequent appearance in the show room. Think here of Maxwell's aether models which facilitated the formulation of his field equations. On the other hand, models can also have a constitutive presence in the theory it serves. Cases in point are the mathematical models for natural selection and the models for general relativity. Constitutive roles endow models with a certain theoretical robustness. They convey unrealities displayed in the theory's showroom. Whatever we might say about such uses, they are on their face both ontically and epistemically consequential. Models carry two strains of unrealism. Models always distort what they model, saying things about them that are false, or omitting things that are true. Even a model that make its motivating data more precise gives them a character they didn't have before.[2] Moreover, as Patrick Suppes points out, the very data of model-based science are themselves really models of them[3]. Since all modelling represents things as they aren't, it is a standing philosophical invitations to embrace scientific unrealism, especially the robust ones. This holds true in whatever reach of science's broad palette robust models achieve a work-permit.

[1] Following Dummett, the preferred expression is "anti-realism", which seems to me to an unfortunate choice. For one thing, these days "anti-" has connotations of partisan nastiness. For another, if it is a less specific term than its better − "unrealism" − which denotes the positive thesis that e.g. mathematical objects– all or some – are unreal objects. In previous writings, I tried to promote "irrealism" as a preferable name, but have now lost confidence in it, in part because of its phonetic and orthographic resemblances to "surreal", which is surely going too far.

[2] See, for example, Woods and Rosales (2011).

[3] Suppes (1962).

When a scientific unrealist drops the word "fiction", more often than not he will mean fiction only in a manner of speaking, to indicate the unreality of its purported denotation. A small minority has something more substantial in mind, and likens the unrealities of its theories to the unrealities of literary fiction. It sometimes adopts the name of fictionalism. A minority of that minority has in mind something both less direct and more determinate. Its boosters explore the possibility of modelling their own fictionalist appreciations of sciences of the unreal in or on some or other theory of the literary unreal. The move is indirect because, instead of modelling the unrealities of science on the unrealities of *stories*, it contemplates modelling its scientific unrealities on how *literary theorists* model the unrealities of stories. But, in as much as there are some more or less well-built literary theories already up and running, it is also a more determinate option to explore.

The purpose of this essay is to weigh the fate of this kind of theory-to-theory modelling. As I hope to show, there is no convincing philosophical good for unrealist science that comes from it, although possibly a power of good for literary semantics.

2 Fictionalism

Fictionalism is a form of unrealism, which in turn is a philosophical theory about scientific theories satisfying the following conditions:

Philosophical Unrealism About Science

(a) Unrealist science convey information designed to advance or close the cognitive agendas of its practitioners.
(b) Some of these information-conveying devices lie in the domains and counter-domains of premiss-conclusion relations implicated in inferences correctly drawn by their users.
(c) Some of the sentences indispensable to the theories' scientific success range over items of no reality, i.e., items which have no real existence.[4]

Some scientists take this philosophical stance towards their own discipline. But when they do, they don't think of using the statement "Numbers are unreal" as a working premiss of physical chemistry, econometrics or the arithmetic of the large cardinals. All that this means is that the working scientist is as free as anyone else is to entertain a philosophical opinion about the work that he does. But the place to look for well-developed explorations of that kind of opinion isn't science, but rather the philosophy of it. Fictionalism is best seen as a species of scientific unrealism which tries in some slight way to compensate for what the entities of unrealist science really *aren't* (namely real) by finding something that they really *are* (namely fictions). We could fairly wonder in what the hoped-for compensation would consist, and how anything unreal could have the property of being something.

[4] In recent years, fictionalism has migrated to various field in which objectivity is thought to be in some peril, notably moral philosophy, philosophy of art, philosophy of law, and metaphysics. Variations of the above three conditions are frequently invoked.

The objects of fictionalism appear to be well-managed in well-respected disciplines that quantify over the nonexistent. Not every practitioner of such disciplines need take the fictionalist view of them, but fictionalists of all stripes assuredly do. It bears repeating, even so, that some fictionalist unrealists drop the name "fiction" as a synonym of "nonexistent" and take no further steps to explicate that equivalence or demonstrate the good of calling on it. Others have something more in mind. They see some value in noting that frictionless planes, infinitely large populations and ZF sets are, like Sherlock Holmes and Baron Charlus, are entirely made-up entities. There are two streams of mathematical unrealism. Leopold Kronecker is famous for having said "God made the integers; all else is the work of man".[5] Kronecker's unrealism was therefore *partial,* as is neoclassical economics in its trafficking with the infinite divisibility of subjective utilities such as pleasure, as in turn is the preoccupation of physics with perfectly rigid rods.

The other strain of unrealism is *total*, grounded in the conviction that the abstractness of its objects precludes their reality.[6] "Fictional" is a plausible tag for made-up science. It would be suitable for totalists who think that all of mathematics is man-made, that is, that its domain is populated entirely by *objets d'art* rather than *objets trouvés*. It is also useful to note the trickiness of the link between an object's abstractness and its impalpability. Abstractness may be sufficient for impalpability, but is not a necessary condition of it. In our relation to him, and his to us Sherlock is an impalpable object, but it stretches credulity that he is in any way an abstract one. This is vigorously disputed by some literary semanticists. They equate Sherlock's abstractness with his ontic incompleteness, asserting that he has only those properties explicitly attributed by the story or follow from them by immediate inference. This is not the abstractness that unrealists attribute to numbers. For one thing, the properties directly attributed to Sherlock are properties of concreta - living in London, being two-legged and pipe-smoking.

By no means are all philosophers of model-based science much tempted by fictionalism, even after some initial brief flirtation with it.[7] Those who find it interesting enough to dwell upon subscribe to three general assumptions.

Philosophical Fictionalism About Science

(i) The subject-matters of fictionalist science are susceptible of truth-evaluable assertion.
(ii) When such sentences occur, they retain their semantic face-value. They wear their meanings on their sleeves.
(iii) The theoretical claims advanced by fictionalist science aren't bound by the standard of truth but, when successful, answer to different standards of cognitive virtue.

[5] Quoted from Gray (2008; p. 19).

[6] The thinking here is that since the number 3^{10} has no spatiotemporal position, it can hardly exist if there is nowhere at all where it ever is. Hardly a convincing argument, it is best to take it as shorthand for the doctrine that there is no branch of natural science capable of explaining either 3^{10}'s eternality or its coming to be as an independent entity – independent of its having been thought up by our forebears. To philosophers who play by the strict rules of platonism, that would be a question-begging argument.

[7] See, for example, Burgess (2010). See also Chuang (2016). On thinking it over, Lui recommends "that one stay away from fictionalism and embrace realism directly".

It is said that modern fictionalism emerged from two different sources in 1980.[8] In *Science Without Numbers,* Hartry Field argues that because of its commitment to the existence of abstract objects interpreted "at face value", mathematics is false. But there is merit in retaining its abstractions for their assistance in regulating our scientific inferences about concreta.[9] In *The Scientific Image,* Bas van Fraassen allows that scientific theories are "genuine interpretations" of unobservables in nature, but their purpose is to play an indispensable role in generating the observability-sentences on which the theories' empirical adequacy rests. Empirical adequacy calls van Fraassen's shots in science; truth is not its objective.[10] The first source is a species of nominalism, and the second the spawning ground of constructive empiricism.

Field's nominalism gives false science a restorative inferential role to play in arriving at the truth, and van Fraassen's constructive empiricism gives the unobservable parts of it a benign role in arriving at empirical adequacy, never mind its truth. Otávio Bueno picks up on this and offers a brand of fictionalism in which the language of mathematics – set theory, say – is left to speak for itself and, when furnished with an existence-predicate, gives us the means of recognizing the things whose existence is not in doubt. However, it takes an agnostic stand toward the existence of those others in the theory's quantificational domain.[11] Conditions sufficient for Bueno's existence-predicate are these: There is a "robust" form of our access to its subject. It enables us to follow it in spacetime in such a way that, had the object not been present to us, we wouldn't believe that it was. (218–219) Bueno's fictionalism is a rival of nominalism, although in some ways sympathetic to it. "Fictionalism", he says, "is here understood as a form of agnosticism: whether mathematical objects exist or not is simply left open." (p. 218).

Moreover,

> "[g]iven its agnostic character, fictionalism is importantly different ontologically from nominalism …. In the end, although it may not be possible to nominalize set theory, it is at least possible to fictionalize it!" (p. 221).

In view here are but three forms of fictionalism, among the several more catalogued in the Kalderon volume alone. It is an abundance that pulls in two directions at once. On the one hand, it suggests that fictionalism about the unreal or the otherwise ontically bothersome is an idea that's ready for serious theoretical development. On the other, it suggests that at this state of its evolution, we haven't quite arrived at the point at which a suitably unified, consistent and stable consensus about its character and its provisions has been achieved. It is right to be mindful of the sharp differences between the claims which attract the unrealist interpretations of the three fictionalisms in play here:

[8] Kalderon (2005). Depending on what "modern" means, this might overlook Leibniz's claim that the inconsistently described infinitesmals were mere fictions and laid no claim to mathematical reality. Closer to now is Vaihinger (1924). In between is Bentham (1932). Vaihinger and Bentham are in Kalderon's bibliography. Leibniz doesn't make the cut.

[9] Field (1980).

[10] van Fraassen (1980).

[11] Bueno (2010).

(α) the claims are false but instrumentally valuable; (β) questions of truth and falsity don't arise for them; and (γ) they are false, and helpful, but our commitment to their purported referents is epistemically underdetermined.

There are plentiful examples in the philosophical literatures in which fictionalism is advanced without reference to the fictions of literature.[12] Of the twelve essays in the Kalderon volume, seven mention literary fiction in their bibliographies, and five do not. None of these seven mention stories, but rather philosophical theories of them. At a certain level of generality, fictionalism models scientific theories of the unreal on literary theories of the nonexistent objects and events of stories, in some hoped-for theory-to-theory mapping. Undoubtedly, literary fictions are in fictionalism's air space, but their effect there varies considerably. For the most part, their touch is light, an idea more toyed with than hitched to dedicated research programmes. This re-prompts our question: What would it take to model the nonexistents of unrealist science on the nonexistents of literary theory, and to what good end? What are the truth conditions of such claims? What, for that matter, are the truth conditions of the fictionalist's assurances that Sherlock never existed and is only a fictional character?

This would be a good place to give a more formal notice to one of the features that dogs fictionalist approaches to the unrealities of robustly modelled science:

The made but not there thesis: Although the unrealities of robustly modelled theories are all made up, they aren't any of them quite all there.

The unrealism of a scientific theory can migrate to any other scientific theory in which it's given some honest work to do, and is also widely present in theories that traffic in idealizations - the infinite size of populations, zero time correlations, the utterly perfect rationality of the utterly ideal reasoner, total information, perfect competition, and so on.[13] This reminds us of an important ambiguity. Sometimes abstractness is equated with non-spatiotemporality (e.g. the number 3^{10} again). Sometimes it is equated with instantiations of comparative properties at some naturally unrealizable limit of instantiation (e.g. perfect competition). The difference between idealized entities and abstract entities is itself subject to confusion. If "abstract" means "unrealized" in spacetime, all the idealizations I've mentioned here are abstract. If an abstract entity is a real entity minus some of its properties, an abstract object needn't be idealized or unrealized in spacetime. Think here of the maquette of the new gazebo for the side-lawn of the tennis club. Euclidean triangles are abstract in both senses.

For some philosophers of science and many of science's practitioners, unrealism carries a whiff of scandal, and even those unrealist faithful who refuse to be scandalized by it are sometimes a bit touchy that something as massively successful and respected as well-made science should be subject to such a philosophically persistent foundational unsettledness. It makes for, as we might say, a certain ontic awkwardness.[14] It is

[12] Quinton (1957).

[13] There, of course, are more varieties of unrealism in science than we could shake a stick at, and a hefty plurality also in the philosophy of mathematics. The best 72 column-linesworth about the realist/unrealist dispute in mathematics written in English are to be found in Blackburn (2005).

[14] And concomitantly, alethic, doxastic, inferential and epistemic awkwardness. Let the ontic stand in for them all.

an awkwardness foretold by the made-but-not-there thesis. It is easy, therefore, to see how it might occur to an awkward unrealist to be on the lookout elsewhere for prosperous unrealisms that carry no hint of scandal and are no occasion for ontic awkwardness or any other. This is not, however, the impulse of most unrealists to date, even those who drop the name "fiction" and those others who espouse fictionalism. Nevertheless, it might be a possibility worth looking into.

The unrealism that lies closest to home for all human beings and is their life-long companion, is the unrealism of stories. It is an empirically discernible fact of lived literary experience that it would never occur to an avid reader of Sir Arthur Conan Doyle that, because of the unreality of their objects and events, his stories are in any way suspect, shady or off-colour, and not quite fit as reading material.[15] This discloses an important difference between scientific unrealists and readers of fiction (and authors, too). The made-but-not-there thesis holds in both cases. But *some* scientific unrealists are made awkward by it, whereas *no* reader of stories ever is. In a way, this is explicable. It is true that all of Doyle's people are made-up, and it is true that they're not there; I mean not *quite*. They are certainly there in the stories, and certainly not there in the world. Theoretical artefacts are different. Artefactualist unrealists know that they're all made up; but they also *sense* that there is nowhere for them to *be*. Readers experience no embarrassment in knowing that Sherlock lived in Baker Street in the 1880s and yet also knowing that he did no such thing. That is not how we respond to perfectly rigid rods. They're not in the world, and if they are invoked, they are in the domain of the theory only in a manner of speaking. Even putting it that way induces a twinge of awkwardness.

Some philosophers have scorned any hook-up with literary fiction. Literary fictions, they say, are trifles, the subjects of our recreational amusements when we've found time to pause from our engagements with truly important things - quantum computing, the 5G race, that sort of thing. Frege did grudge that there is *some* place in human experience for "Odysseus was put ashore at Ithaca", but he insisted that it possessed no scientific interest. By this he meant that it held no importance for what interested him, to wit, the foundations of arithmetic.

One prominent strain of mathematical unrealism ensues from the inconsistency of intuitive set theory, established by Russell in 1902 in a proof that fuelled Frege's abandonment of logicism. It is the proof that there exists a set that concurrently and unambiguously is and is not a member of itself. It promoted the effort a year later to stipulate consistent successor-entities by what Russell called nominal definition and what Quine later on would call myth-making.[16] If the post-paradox sets that were

[15] Irrespective, of course, of their artistic merit, or their serving as distractions from more pressing business emanating from head office.

[16] Russell (1967). Russell (1937). Quine (1966; p. 27). I mention in passing that Russell's adoption of logical fictions is not fictionalism. Russell's logical fictions are logical constructions, as numbers are of sets. Russell never thought that the reduction of numbers to sets established their nonexistence. So logical fictions aren't fictions.

made-up for the natural numbers were objects of myth-making, how much less so could the sets that were made up for the inaccessible cardinals be?[17]

In some quarters, it is casually put about than an enlarged understanding of how scientific theories of nonexistent entities manage to enlist the doxastic support of serious and well-informed people might be obtained by modelling the scientific theories in (or on) "our best" theories of stories. But in the absence of a detailed command of the data that motivate story-theorists to provide a cohesive and unified account of them, any such modelling aspiration is a risk-laden prospect for any science that countenances the unreal. Below I shall bring to light some perhaps startling peculiarities of stories which, to the best of my knowledge, have yet to find a seat in "our best" science of literary fiction. Once the problems they provoke are solved, we will be in a better position to assess the model-worthiness of unrealist science in a successful theory of stories. Before going there, we should spend some time with a philosophical trainwreck that lies in wait for scientific unrealisms of every kind.

3 Grundgesetze der Semantik

Given the prevailing logico-semantic standards of our day, a theory's unreal objects cannot be the values of the bound variables of its quantifiers and can't be the nominata of its nominal names or referents of its other singular terms. If the arithmetic of the inaccessible cardinals is about nothing whatever, it is hard to see anything interesting or instructive in it. At the centre of this puzzle are Basic Laws - *Grundgesetze* - broadly taken by contemporary philosophers of language as regulating the logico-semantic traffic in human speech and thought. They come straight out of the model-theoretic formal semantics for classical first-order logic.

I *The something law*: Everything whatever is something or other.
II *The existence law*: Reference and quantification are existentially loaded.
III *The truth law*: No statement-expressing sentence that discomplies with either of the above laws can be true.

Anyone who acknowledges Sherlock's nonexistence and accepts the three laws is committed to accept a fourth:

[17] A cardinal κ is (strongly) inaccessible if it is uncountable, not a sum of fewer than κ cardinals that are less than κ and $\alpha < \kappa$ entails $2^{\alpha} < \kappa$,. Informally, an uncountable cardinal is inaccessible if it cannot be obtained from smaller cardinals by means of the standard operations of cardinal arithmetic. See Sierpinski and Tarski (1930). See also Zermelo (1930). Especially recommended is Kanamori (2001).

IV: *The fiction law*: Sentences of fiction fail to refer and cannot be true or false.[18]

It will be instructive to log some examples of what these laws preclude. The something law strikes down "Some things are in no sense anything at all; they are nothing whatever." The existence law disbars "There are things that don't exist" and "Sherlock Holmes is merely a fictional character". The truth law overturns "Vulcan was hypothesized by Jacques Babinet to be the planet that reconciled the orbital perturbations of Mercury with Newton's laws." With but two exceptions, I have been unable to find in the western semantic canon any resistance to Basic Law I - the something law - or a touch of regret about its exclusion of "Some things are nothing whatever - complete and utter nothings".[19] Meinong's objects of which it is true that there are no such objects, i.e., that they are nothing at all, seems to be one of the exceptions.[20] More recently, we have Graham Priest's nonexistent objects which have individuating properties, that separate one object from the other, yet are absolutely nothing at all, not even apparently the bearers of those differentiating properties.[21] There is a favourable review of *Towards Non-Being* in the *Times Literary Supplement* of February 23, 2018. But not even its reviewer, Tom Graham, is able to absorb the idea of a nonexistent object's capacity for having honest-to-goodness properties yet being absolutely unable to be anything at all. The reviewer rejects Priest's advice to puzzled readers to "get their understanding rewired." I reject it too. For the present I'll take it as read that the exceptions are unavailing and that the something law is safe as churches, which is the last thing existence and truth laws are.

Riding along with the semantic and epistemic vacuity of literary discourse about the unreal is a puzzle about how to account for the enormous interest stories attract and the high regard in which we hold their makers. It is, to be sure, a real problem, not merely an annoyance. Since, as proponents of the Basic Laws aver, Sherlock doesn't exist, he can't be named or otherwise referred to. If he can't be referred to, nothing can be

[18] It is easy to see the lineaments of these four in the metalogical wherewithal of classical logic, in which the existence predicate gives way to quantifiers that take on the connotation of the vanished predicate. The existence of x is now conveyed by "$\exists y\ (y = a)$". The link between the something law (I) and the existence law (II) is supplied by the EG proof rule: If a is F, then there exists an x which Fs. The trouble caused by the Basic Laws is sometimes said to arise from Mill's philosophy of language. I am unable to see why. Mill himself writes that "[a]ll names are names of something, real or imaginary." (J. S. Mill, *A System of Logic,* 3rd edition, London: Parker, 1843; p. 27.) Yet Sainsbury writes, "At one point I thought that it was clear that [names of the fictional] could be used as examples of names without bearers, and so as counterexamples to Mill: *they have no bearer,* so on Mill's theory have no meaning; but they are plainly meaningful, so Mill's theory must be abandoned." (Sainsbury (2010; p. xviii); emphasis mine.) Later on, Sainsbury mentions the dismissal, by van Inwagen, of attributions of violin-playing to what Sainsbury characterizes as "the abstract artefact Sherlock Holmes", which van Inwagen calls "a silly mistake". (Peter van Inwagen "Creatures of fiction", *American Philosophical Quarterly,* 14 (1977), 299–308; p. 306). This is not the place to dwell on these Millian confusions. For our purposes, it suffices to lodge the Basic Laws of Fiction in the metalogical provisions of classical logic, adjusted for natural languages in canonical notation.

[19] Except as an insult.

[20] Jacquette (1996; p. 9).

[21] Priest (2016).

ascribed to him, e.g. that he was a resident of Baker Street. If nothing can be ascribed him, nothing can be said about him (including, paradoxically, by the very sentences I'm now penning). If nothing can be said about him, nothing true can be said about him, and nothing false either; and Sherlock would have no readers. *None.* No story would.[22]

The same holds, line for line, for the natural numbers, ZF sets, frictionless planes and infinitely divisible utilities.[23] If the logico-semantic orthodoxy were to prevail, unrealist science couldn't be interesting. It couldn't be understood. We couldn't know what it says. It couldn't be believed. It would give us no knowledge of anything and would provide no occasion for the bestowal of a Field's Medal or a Nobel Prize, clear motivation for semantically fresher air in which to do the business of science.

4 Fresh Meinongian Breezes

Fortunately, the present-day formal ontology literature is dignified by a distinguished minority of theories which retain the something law - Basic Law I - but strike down Laws II and III, the existence and truth laws. When they extend themselves to fiction, they also strike down the fiction law IV. They are all variations on the theme of Meinong's ontology of objects.[24] Meinong himself was not preoccupied with the logic of literary fiction, and played no role in the emergence of fictionalisms of the sort under review here. His was a psychologically informed exercise in ontology, in which a wholly general theory of objects in their various kinds would be worked out. Modern Meinongian logicians seek to formalize the main features of Meinong's ontology in the same kind of way that mainstream logicians deal with truth and consequence in the model theory of classical logic, in the manner of Tarski for example.[25] The key differences between the classical approach and that of Meinongian logicians is that

[22] One establishment-approved manoeuvre to save the day originates in Russell's "On denoting", *Mind* N. S., 14 (1905), 479–493. Reprinted in Alasdair Urquhart, editor, *The Collected Papers of Bertrand Russell*, volume 4, *The Foundations of Logic, 1903–05*, pages 415–427, London and New York: Routledge, 1994. The method known as contextual elimination would reconstruct sentences such as "The most famous deducer in England resided at 221 Baker Street in 1880 s London" in a way that eliminated the referring term and drew upon existential quantification to produce the false sentence "There is one and only one object who is the most famous deducer in England, and that very object resided at 221 Baker Street in the 1880 s." Later, names in referring position were subject to the same displacement and their contents were repackaged as predicates – "socratizes", "sevillaizes", and so on. "Sevilla is lovely" would be recast as "There is one and only one thing that sevillaizes and that very thing is lovely." What one loses here is not just the non-philosophers' universally contravening belief that it is true that Sherlock lived there. Of more immediate importance, aboutness is systematically suppressed in fictional contexts. So, again, the sentences in question says nothing about anything, and can't say anything at all that implies or presupposes their aboutness.

[23] It is worth noting that the backers of the Basic Laws take them to apply universally, that is, to any mode of discourse or theoretical investigation whose subject-matters, in whole or part, are unreal.

[24] Meinong (1904); Routley (1981); Parsons (1980); Jacquette (1996); Berto (2013); Berto and Plebani (2015).

[25] Tarski (1983).

classical logic is steadfastly extensional and Meinongian logic is steadfastly intensional. Classical logic puts the squeeze on aboutness, whereas the Meinongian alternative gives it free rein. Classical aboutness is squeezed to the point of breathlessness by the Basic Laws. Meinongian semantics breaks the aboutness-stranglehold.[26]

In some respects, Meinongians escape further from the pinching effects of Law abiding semantics more than is altogether good for fiction. It provides that when we say "The golden mountain exists", there is something determinate that we refer to. If we consulted the empirically discernible data of lived referential experience, we'd see that we experience ourselves as doing no such thing. Quite apart from an object of which it is true that nothing whatever is it, this is an ontology surplus to need in our understanding of human cognitive practice. What is more, Meinongeans who endorse the restricted comprehension principle, load up their ontologies with masses of incomplete objects, as follows:

> *The restricted comprehension principle*: For any nuclear condition F*x* with free *x*, there is an object that satisfies it.

Roughly speaking, an object's nuclear properties are the properties that constitute it – e.g. the property of being six feet tall – and its nonnuclear properties are properties it has but is not constituted by – e.g. the property of being a possible thing.[27] The risk to Sherlock and his like is that if his nuclear or constitutive properties are exactly those ascribed by the sentences of Doyle's text, he and they would be both ontically incomplete and readerless. They would be metaphysical *freaks.*[28]

---

[26] It is important that we not confuse intentions with intensions. They are fraught matters in analytic philosophy, but they are different issues. A Meinongian intention is an object of thought; it is what the thought is a thought *of.* It is an "intended" object, the object towards which the thought is directed. Similarly for statement-making sentences. They, too, are directed towards the things they are about. Suppose I am now thinking of all the things that are cordates and later am thinking of all the renates. I am each time thinking about the very same things, those who have hearts and also have kidneys. (The example is Frege's.) Intentionalists agree that the property of being a cordate and being a renate have. Some philosophers, notably Quine, distrust all talk of properties, as opposed to sets of their purported instantiations. Quine's complaint is that, whereas sets are well-individuated objects, properties are no such thing, calling into question whether the property of being a cordate actually qualifies for what any thought or sentence could be about. I lack the space to linger with this, but not before remarking that no set is well-individuated unless its members are. Think here of the set of all the individual clouds in Sevilla in the cloudiest period of its calendar in 2019. Right or wrong, Quine's point is that in the absence of successful conditions of individuation, the thought and the statement-making sentence can't be assured of referents, hence can't be assured of there being something they're about. Intensions, therefore, are a menace of intentions. In a variation, intensions are said to be meanings, and draw the same extensionalist complaint. Relatedly, what makes a logic intensional is its tolerance of exceptions to the uniform intersubstitutivity law of classical logic – e.g. "On St. Valentine's Day John is thinking about his favourite cordate" need not hold up if "renate" is substituted for "cordate". For the intensional and extensional, see Zalta (1988), as well as Jacquette (1996). For the anti-intensionalist pushback, see Quine (1986). For more on aboutness, see Yablo (2014a), and Simchen (2017).

[27] See here Jacquette, *Meinongian Logic,* section 5.1.

[28] For further reservations about how the nuclear-nonnuclear distinction misfits the facts of lived literary experience, see my *Truth in Fiction: Rethinking its Logic,* volume 391 in the *Synthese* Library, Dordrecht: Springer, 2018, chapter 8 "Other Things Sherlock Isn't", section 8.2.

There is one point on which classical and Meinongian logicians are agreed. They agree about methodology. Imagine some subject-matter S about which we would like some theoretical elucidation. How should we go about achieving an improved understanding of the S- data? The answer in each case - Meinongian and classical - would be to formalize them by constructing a model theoretic semantics.[29] Formal models of each persuasion are set-theoretic structures, leaving the inference that, for the things of interest here, enlarged theoretical understandings are to be found in mathematics.[30] This is a show-stopping development. Fictionalists who seek relief in the formal semantics theories of literary fiction seek it for the relief of their own scientific unrealism. Sets, in their turn, are prime candidates for unrealist construal. The semantics that formalizes literary discourse is a theory that models it mathematically. To be clear, I am not claiming that Meinongians themselves seek relief anywhere from the irritations of Meinong's own unrealism. Meinongian ontologists aren't worried about the unrealities of science, and Meinongian semanticists are happy to preserve them. What I am saying is that anyone who's worried about the unreality of sets cannot possibly assuage it in a set-riddled formal semantics, classically formatted or Meinongianly so. It is a footless endeavour, on a par with trying to escape the winter blasts of Eureka in the Canadian territory of Nunavut by seeking relief in Karosjok, in Norway, rather than Sevilla in Spain.[31] This is a serious enough blunder to have a name. So let's call it

The Eureka to Karosjok winter-relief blunder

This also gives us occasion to consider the attractiveness of a quite different approach. It would favour theories whose motivating data are empirically discernible facts of normatively assessable cognitive behaviour on the ground, without resorting *ab initio* to the impulse to mathematicize them. The logic that does best for such data could turn out to be one that accords full standing to those empirical facts, subject only to the condition that it should take formidable cause to ignore or contradict them. This would be a *naturalized logic,* circumspectly joining forces with the psychology, sociology of knowledge and the other branches of cognitive science, AI and neurobiology.[32]

[29] I should disclose my own surrender to the gravitational tug of formal semantics in *The Logic of Fiction: A Philosophical Sounding of Deviant Logic* (The Hague and Paris: Mouton, 1974); 2nd edition, with a Foreword by Nicholas Griffin, volume 28 of Studies in Logic, London: College Publications, 2009. My preference then was for a quantified modal semantics for fiction.

[30] Concerning which, see Aristotle: "In general, though philosophy seeks the cause of perceptible things, we have given this up … [for] mathematics has come to be the whole of philosophy for modern thinkers." (*Metaphysics,* 992[a] 24–31.) Of course, this is a view that Aristotle himself steadfastly rejected. See also the admonitions of Kant's U*ntersuchungen über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral,* Munich: GRIM Verlag, 2007; first published in 1764. These alert us to the potential of bad ideas to persist. For more push-back, readers could consult my "Epistemology mathematicized", *Informal Logic,* 33 (2013), 292–331.

[31] The mean February temperature is −55 °C in Eureka and −51.4 °C in Karosjok. It is 12.5 °C in Sevilla, a rounded sixty degrees warmer than Eureka.

[32] Further details can be found in my *Errors of Reasoning: Naturalizing the Logic of Inference,* volume 45 of Studies in Logic, London: College Publications, 2013; reprinted with corrections in 2014. See also *Truth in Fiction,* chapter 2 *et passim.*

5 Naturalized Logic

I come now to the subject-matter of my subtitle. It might strike some readers as an unnecessary distraction, for isn't scientific fictionalism our subject here? Yes, but a larger part of it is the semantics of literary fiction on which some versions of scientific fictionalism seek to model themselves. As I propose to show, we aren't going to get literary semantics right without taking a naturalistic turn. A naturalized logic is a theory of cognitive practices that are empirically discernible and subject to participants' judgements of well or badly done. It carries further procedural desiderata. One is that data of material relevance not be overlooked. Another is that data not be misconstrued, that is, not be subject to tendentious data-bending in quest of some pre-determined theoretical outcome.[33] The theory should avoid errors or omissions of data-collection and data-analysis. We could call these collectively *the respect for data principle*. Also important is the theory's empirical sensitivity; by which I mean that when one's own theory appears to transgress some lawlike regularity of a well-confirmed and duly replicated empirical theory of material relevance to it, one should not disregard it save for well-considered cause to override it. This is *the empirical sensitivity principle.*[34] Something also to aim for is that, as much as possible, we keep antecedent philosophical preconceptions in proper check. This is especially important if the preconceptions are motivated by different (kinds of) data from the present ones – e.g., when the different data are concepts, semantic intuitions and the like. Call this *the constancy of data or apples-to-apples requirement*, and add it to the respect for data principle. A final desideratum is to set out naturalized criteria for normative adequacy. This is *the naturalized normativity condition.*

For ease of reference, we can call all these desiderata *the naturalistic constraints* on a logic of real-life cognitive behaviour. A logic that disobliges the naturalistic constraints misses the mark by not with all requisite circumspection hitching its wagon, to the best of our naturalized versions of epistemology. The one that works best for realities of lived epistemic experience is a strongly causalized adaptation of reliabilism. To give a scent of it, here is a simplified formulation of how it provides for someone's knowing something.

> *Causal response knowledge*: Agent X knows at *t* that *p* on information I when *p* is true at *t*, in processing I belief-forming devices causally induced X to produce the belief that *p*; X's devices are in good working order and operating here as they should; I is good and properly filtered information, and there is no interference caused by negative externalities.[35]

[33] Gigerenzer (1996).

[34] It is prudent to keep it in mind that prominent empirical theories sometimes employ modelling methods that are insufficiently explained or which over-distort their motivating data. In some respects the value of the empirical sensitivity principle is maximized by attending to the theories' *data* rather than their own theoretical provisions for them.

[35] The closest view to this one is to be found in Alvin Goldman's first paper on the subject, which I take to be the first modern step into what Quine shortly later would call "epistemology naturalized". See Goldman (1967). The principal difference between Goldman's approach and this one, is that Goldman causalizes the justification condition, whereas I deny it as necessary for knowledge in general, and apply the causal conditions directly to knowing.

Let's quickly add that information is good when it is accurate and up-to-date, and is well-filtered when it is unmolested by noise, by quantities that overload processing capacity, by distortive irrelevancies, by unproductive, by wheel-spinning inconsistencies or by the smothering ignorance. Information-processing is impaired by the smothering ignorance-effect when, in processing it, an agent is over-mindful of the massiveness of what's not known to him.[36] The nub of it all is that information-processing often produces belief which, when true, places the processor in a state of knowledge. Knowledge, then, is *well-produced* true belief. The larger discussion of this idea can be found in E*rrors of Reasoning*. I mention it here to give some hint of what an epistemologically sympathetic treatment of our cognitive intercourse with fiction would look like.

One of the dominant regularities fiction-reading is that, when we read the stories, we are moved to believe the things they tell, and are brought to a state of knowledge about what goes on there. This is a datum of utterly foundational interest – and completely at variance with the Basic Laws. Here too, the respect for data principle kicks in. We should attend to the ways and means of our doxastic and epistemic responses when we read the stories that engage us. But we shouldn't set out to do this without turning our minds to empirically discernible details about how we, the beings who have those experiences, are cognitively structured, and set up by nature and nurture to know things. So, then, every living soul, including the smartest of our species, takes it as given that:

The cognitive abundance thesis: Beings like us know lots and lots of things about lots and lots of quite different things. We are successful and versatile knowers.

The error abundance thesis: Beings like us make lots and lots of errors about lots and lots of different things. We also have feedback mechanisms that provide for the detection and correction of error. In the general case, this is a more efficient means of maintaining cognitive equilibrium than error-avoidance no matter what.

The enough already enough thesis (Hereafter "Enough"): Although we are prone to error, we are right enough of the time about enough of the right-enough things to survive, prosper and, from time to time, build great civilizations.

We could call these the Three Pillars of a credible epistemology of human knowledge. For the three to be true together, we must have a Fourth Pillar.

Error-management: There is a considerable economic advantage in the efficiencies of feedback loops that enable the timely detection and high-frequency investment in error-avoidance up front.

The Enough thesis is something to dine out on. It reveals the human agent as a dab hand at responding to cues in cognitively productive ways. It reveals him as a being well-versed in the acquisition of knowledge and the mitigation of error. It casts pre-emptive doubt on any theory of fiction that discounts the lived realities of literary

[36] Further details can be found in my "Four grades of ignorance and how they nourish the cognitive economy", Synthese, https://doi.org/10.1007/s11229-019-02283-w.

engagement. It puts on notice any silly notion that when it comes to their engagements with stories, readers and writers the world over have fallen off their cognitive perches have simply lost their minds.

We now have the means to comply with naturalized normativity condition. For the Four Pillars to be true, we must also have

> *The N-N convergence presumption*: In the general run of things in human life, good reasoning on occasions K is reasoning as it normally play out on K-occasions. Presumptions, of course, are defeasible. They are normally invoked in the expectation that the burden of proof rests on purported exceptions. It is rightly considered a meetable standard for fair cause.[37]

As we see, a formal semantics lies at least one more remove from these empirical data than where a naturalized logic of them lies, and requires the appropriate formal representability proofs (hardly ever attempted) to re-establish credible contact with them.[38] In these matters of normatively assessable human cognitive behaviour, I am logically more drawn to the natural than to the mathematicized formal, notwithstanding the technical virtuosities involved in going mathematical.

It is easy to see that a good part of the explanation of our cognitive serenity about fiction's systematic inconsistencies lies in an adoption of a Meinongian approach, *minus* Meinong's theory of objects, and *minus* Meinongian semanticists' fondness for mathematical models, from which we would have

- the repeal of Laws II, III and IV

  and

- the retention of a freed-up aboutness relation, as essential for explaining human cognitive behaviour.

[37] I'll come back to the normativity question in section 8.

[38] I discuss these data-to-theory gaps in greater detail in "Does changing the subject from A to B really enlarge our understanding of A?", *Logic Journal of the IGPL,* 24 (2016), 456–480. For a penetrating discussion of the inadequacies of mapping and quasi-mapping approaches to formal representability, see Robert W. Batterman, "On the explanatory role of mathematics in empirical science", *British Journal for the Philosophy of Science,* 61 (2009), 1–25. For the mapping approach see, for example, Pincock (2012); Bueno and Colyvan (2011). Something else to bear in mind rather more generally are the current difficulties concerning the Equilibrium Climate Sensitivity (ECS) measure of how the climate responds to greenhouse gases. There are currently two inequivalent ways to estimate ECS. One is the model-based method. It is a method that generates the prediction that ECS is probably between 2 and 4.5 °C, possibly as low as 1.5 °C but not lower, and possibly as high as 9 °C. It is a significant prediction range. If the ECS is around one, climate change is not worth bothering with. If it clocks in at four or more, it is enough of a problem to make it questionable whether there is anything to be done to fix it. The present method begins with a climate model. The other method, the energy balance method, begins with historical data – long-term historical data on temperatures, solar activity, carbon-dioxide emissions and atmospheric chemistry to estimate ECS using a simple statistical model derived by applying the law of conversation of energy to the Earth's atmosphere. The energy balance measures are much lower than the model-based measures. In a recent formulation, its estimate is 1.5 °C, with a probability range between 1.05 and 2.45 °C. This is not the place to take this further. It suffices to say that in the model-based business, it really does matter what one starts with. See here Dayaratan et al. (2017).

When we add a third nugget,

- a determination to respect the empirically discernible facts of human referential ascriptive, alethic, doxastic, cognitive and inferential behaviour in the actual conditions of real life,

and to it a fourth, which lies at the heart of open-space aboutness,

- an affirmation of real relations between palpable objects such as ourselves and impalpable ones such as Sherlock, the number 3^{10}, and infinitely large populations,

we would have a promising shot at a stable logico-semantics for stories.[39] We would have profitable occasion to say things about Sherlock, engage with him in ways that enable true beliefs and real knowledge of him, and have coherent and truth-evaluable discussions with others about him and his doings. If, with an appropriate tightness and an absence of anomalous outliers, unrealist science were to snug up to this view of things, its sponsors could say their piece without awkwardness about their own favoured unrealities.

All the same, it would be wise not to overlook the possibly discouraging asymmetries between the real but impalpable relations we bear to the nonexistents of fiction and the real but impalpable relations we bear to the nonexistents of unrealist science. In the case of Sherlock, there is a clear explanation, open to empirical inspection of the facts of lived literary experience, about what initially brought to pass our real but impalpable relations to Sherlock. There is no like explanation of how these same relations to large cardinals, perfectly rigid rods and the infinitely sliceable pleasure of a well-made bourbon manhattan were initially brought to pass. We know who Sherlock's creator was and what it took for him to bring us into contact with him. He did it by making things true of him. Population biologists didn't make it *true* - even in population biology - that populations are infinitely large. Classical decision theorists didn't make it *true* - even in classical decision theory - that decision makers close their beliefs under consequence. Yet Doyle *did* make it true that Sherlock in the stories lived in Baker Street. There are truth-makers for the untruths of fiction. The information conveyed by the stories were processed in ways that produced the Baker Street belief, but not the cardinality belief or the divisibility untruth. There are no truth-makers for the untruths of population genetics and neoclassical economics. When functioning properly, stories push the causal buttons of belief. The others do not. This tells us something important about our belief-inducing devices. They are *filtration* devices, screening out the unbelievable from the believable.

[39] In an interesting reversal, Danny Daniels borrowed the concept of story for the formal semantics of a first order language supplemented by the sentence-operator "the story – says that –". The set Σ is made up of stories *s*. A sentence A is true in a story *s* just in case A ε *s*. A implies B iff for all stories and the actual world, whenever A is true, B is also true. The system is sound and complete with respect to its semantics. The basic language is described in Daniels (1987), and the story-operator appears in "'The story says that' operator in story semantics", *Studia Logica,* 46 (1987), 73–86. The prior paper is announced in this one as appearing under the title "A story semantics for implication". I don't know the reason for this switch. Story semantics, as we see, is a variation of the more standard possible worlds kind.

6 Ambiguity

Let's come back to that arresting peculiarity of our intercourse with the fictional. Readers know perfectly well that none of it is true and yet at the same time know perfectly well that Sherlock lived in Baker Street and that he did not live in Cheyne Walk or Tavistock Square. Or Calle San Fernando.[40] There are some standard manoeuvres for dodging the charge of inconsistency, of which the most popular by far is *ambiguation*. Everyone is familiar with the good that *dis*ambiguation does. We liberate a sentence from apparent falsity or inconsistency by exposing a discernible ambiguity in which the apparent problem resides. The stock examples are bachelors who aren't men and banks that border no bodies of water. Ambiguation works differently. It attributes ambiguities which, if actually present, would indeed solve the problem at hand. In the case of fiction, the candidates for ambiguity-attribution are the predicates "true" and "false", the names and singular terms of fictional items, and the sentences making subject-predicate reference to them. The purportedly resolving ambiguities are either lexical - in the manner of "I want to get to the bank promptly" – are you a swimmer or a depositor? Or they are syntactic, in the manner of "Visiting relatives can be boring" – the visitors or the visited? It is true - and pivotally important - that "Sherlock lived in Baker Street" is true in the story and false in reality. But nothing to be found in theories of meaning for English by empirical linguists lends any field-support to the notion that the in-the-story/in-the-world dichotomy is an ambiguity-marker for "true", "false", "Sherlock" or "Sherlock lived in Baker Street". It is one thing to allege ambiguities. It is quite another to demonstrate them. Kripke is good on this point:

> *Kripke's ambiguation-avoidance principle*: Do not "postulate ambiguities just in order to avoid trouble, for any trouble can be avoided that way."

We'll have to find other ways to explain the cognitive serenities that attend our congress with fiction's systemic inconsistency.

Influenced by the plain fact that "Sherlock lived in Baker Street" is true in the story and false in reality, it is open to fictionalists to invoke a similar distinction for unrealist science: "true in the theory/false in reality." All I will say about this here is to repeat that there is a marked difference between showing that something is true in a story and showing that something very different is true in a scientific model or a scientific theory. Part of the difference, especially among mathematicians, is a readiness to equate truth with provability. There is room for both in fiction but they are not equivalent there. Of course not, how could they be? "True" and "provable" aren't natural language synonyms. The bigger question pertains more to the truth-part of the present suggestion than to the showing-part. There are several excellent accounts of how people came to speak ZF sets into dominance, but nothing close to consensus about its alethic payoff, e.g., about the relationship between objectivity and practice.[41] For the present, I'll stick

[40] Quoted from Berto and Plebani (2015), p. 112. See also *Truth in Fiction,* chapter 6, section 6.3 "Ambiguity".

[41] See, for example, Ferreriós' (2007) and Gabbay et al. (2012). See also Kanamori (2013).

with the assumption that the physical sciences and mathematics are amenable to the same one-size-fits-all fictionalist treatment. In Sect. 11, we'll have occasion to question these assumptions.

As we have seen, an unrealist who thrums to the call of, "Let fiction be your guide!", has two sources to reflect upon. One place to look is the network of worldwide facts of lived literary experience. Another source is the wherewithal of philosophical theories of fictional discourse. As we also see, fictionalism has launched unrealists into the jaw-locking conflict with the mainstream analytic philosophy of language which we've lately been reflecting on. If we compare Basic Laws II, III, IV with the three conditions (α), (β) and (γ) that characterize the unrealisms under review here, we see a fatal discordance. There is also a failure of fit between our (α), (β) and (γ) and the three conditions on the general notion of fictionalism, (a), (b) and (c).[42] No account of fiction that submitted to Basic Laws I to IV could handle the modelling for any unrealist theory that honoured conditions (α), (β) and (γ). Virtually all the going theories of storied-fiction are not theories about the facts of lived literary experience. They are theories that respond favourably to their sponsors' pre-theoretic intuitions about literary semantics which, in every case, is an adaptation of some or other already up-and-running formal semantics for some or other fragment of natural language; or for less formalized ones, such as pretendism (of which, more shortly below). Since the first three Basic Laws have majority standing in those places, it only stands to reason that the majority view of literary semanticists would accept their implication of Basic Law IV, which is the *very* law that gives the hook to the empirically discernible facts of lived literary experience. Among the casualties are the sentences of a story's text, of course; but also sidelined are what *we ourselves* say about their purported objects and events – e.g. "Sherlock is a character of Doyle's creation" and "Dr. Watson was never in India, and no gunfire in any real subcontinental uprising caused him the slightest inconvenience." Thus is any theory of fiction that subscribed to the Basic Laws hoist on its own *pétard*.

7 Pretendism

Kendall Walton's *Mimesis as Make-Believe* is an important case in point. For fictionalists who make mention of literary theories, *Mimesis* is the most cited philosophical example. It is an odd preference. *Mimesis* has the inconvenient feature of enforcing the Basic Laws of Fiction. But Walton doesn't think that the Basic Laws of Fiction in any way imperil what we, who discuss fictions, are up to when we do.

[42] Here they are again: (a) Unrealist theories convey information which advances or closes the cognitive agendas of their makers regarding items of material interest. (b) Some of these information-conveying devices lie in the fields of premiss-conclusion relations implicated in varying ways in inferences correctly drawn by their users. (c) Some of the sentences indispensable to the theories' scientific success range over items of no reality, i.e. items which have no real existence. Compare with: (α) the claims of unrealist science are false but instrumentally helpful. (β) Questions of truth and falsity don't arise for them. (γ) They are false, but our commitment to their purported referents is epistemically underdetermined.

I won't take the time to review in detail this exercise in risk-evasion, except to say that Walton's view is that when we say that Sherlock was a wonderful deducer, we are actually doing something quite different. We are merely *pretending* that he was a wonderful deducer, or in a variation, merely pretending to *say* that he was a wonderful deducer. This raises three difficulties, one taxonomic, another philosophical and the other empirical. The taxonomic problem is that pretendism discomplies with the fictional requirement that the sentences retain semantific face-value. Pretendism, therefore, is fictionalist in name only; it is pretend-fictionalism. The philosophical difficulty is how to parse "We pretend (to say) that Sherlock was a wonderful deducer" if the Basic Laws hold true. Who or what is the object of our pretence? Is it we or is it Sherlock? If nothing whatever is that object, what does "We pretend (to say) that Sherlock was a wonderful deducer" tell us? What are its truth conditions? The empirical difficulty is that no one in the whole wide (non-academic) world would give either version of this fanciful idea the time of day.

It is to Walton's credit that in the early pages of *Mimesis* he makes it clear that the book's ascriptions of pretence is not pretence in the customary uses of that word. What, then, does he mean? What are the technical meanings of his theoretical vocabulary, "make-believe", "pretend", "prop", "play-acting" and so on? Walton has no direct answer, nothing along the lines of "yclept" means "known as". He proposes instead that the technical meanings will reveal themselves just by reading the book to its end. Having read *Mimesis* cover-to-cover several times in the past twenty-eight years, I used to think that the penny had yet to drop for me. I used to think that if Walton's technical vocabulary were indiscernible until the book's ending, it could not have attracted and held readers, or be a favoured nesting place for interested fictionalists.[43] A moment's reflection, which came to me late, puts paid to that misbegotten and empirically false expectation. We must allow that, when they absorb the pages of *Mimesis,* readers act on the implicit meanings of Walton's otherwise opaque text. Given Walton's own warning at the outset, they interpret the book's technical vocabulary as *approximating* to the ordinary meanings of those same expressions. Bradley Armour-Garb and James Woodbridge have made the useful suggestion that Walton's technical vocabulary *precisifies* its ordinary meanings.[44] It is a helpful idea. It helps explain why, from page one onwards, *Mimesis* has had readers, lots of them. But it adds a further complexity to the tale, three of them, actually.

- One is that the precisified meanings are themselves left completely unexplicated.
- Another is that in this unexplicated state, they themselves take on the gravamen of undefined theoretical postulates.
- A third wrinkle is that as the technical meanings more closely approximate to the natural ones, the greater is *Mimesis*'s exposure to empirical discouragement.

[43] See, for example, Frigg (2010). *Mimesis* appeared in 1990 with the Harvard University Press.

[44] The idea arose in a panel discussion of *Pretense and Pathology* at a 2016 Book Symposium at Seattle meeting of the Pacific Division of the APA. In addition to the authors, the panel included Fred Kroon and me. Authors' and critics' comments, supplemented by contributions from Jody Azzouni, now appear in an *Analysis* book symposium, *Analysis,* 78 (2018). My presentation, "Pretendism in name only", appears at pages 713–718.

And, of course, there remains the problem of specifying truth conditions. Whatever the precisified meaning might be of "We pretend (to say) that Sherlock was a wonderful deducer", what possibilities left to us by the Basic Laws? Certainly not that Sherlock is the object of our pretence; certainly not the bearer of the properties we pretend to ascribe to him; and certainly not the subject of the truths we pretend to tell.[45]

Walton's is but one example of attempts by literary semanticists to have their cake and eat it too, pledging fidelity to the Basic Laws while feigning non-offending ways of saying their piece about fiction.[46] It is an interesting manoeuvre in which the theory's logico-semantic respectability is preserved by artful evasion. The trouble is that in cases like Walton's the evasive measures themselves also discomply with empirically discernible facts of lived literary experience.

Before leaving *Mimesis*, it would be wise to re-emphasize that the pretendings, make-believings, play-actings, and so on in play there are *themselves* the theoretical entities of a philosophical account of fiction and are subject, therefore, to the same general latitude and constraint that attends the postulation of theoretical entities in science. This puts Walton's theory on the same footing of those unrealists who look to Walton's for logico-semantic relief, and puts essential pressure on the key question of how Walton views his own theory of the unrealities of fiction. If, as it appears, it is a realistically intended theory, it is almost certainly false. If it is itself an unrealist theory that traffics in the made-up doings of the semantically re-purposed vocabularies of natural language pretence, it loses contact with the empirical realities of humanity's literary experience and, in so doing, puts itself in the very fix that many fictionalists worry that mathematics is in. This, as we see, is the Eureka-to-Karosjok blunder yet again.

8 Normativity

It is not unknown for an empirical failure to seek redemption by invoking the distinction between descriptive accuracy and normative authority. Classical decision theory is a stock example, postulating that decision-makers have perfect information and close their beliefs under logical consequence, each false and in no finite degree approximating to anything in reality. Not even the smartest decision-maker in all of human history makes decisions in that way or aspires to do. Although the classical theory is falsified by the known empirical facts, its sponsors think this a trivial immateriality, made so by the fact that the postulates are normatively authoritative for rationality at its idealized best. If that were actually so, a cognitively competent human being's rationality would be some measurable approximation to this ideal. To the best of my knowledge, no such approximation relation has yet to be adequately defined for

[45] Pretendism, as we may note, is a variation of the ambiguation manoeuvre, in effect purporting the ambiguity of "say", "assert", "ascribe" and so on, with the second meanings of which all taking the prefix "pretend to".

[46] Other pretendists of note include Searle (1975); Kripke (2013); Kroon (1992); Sainsbury, *Fiction and Fiction and Fictionalism;* Yablo (2005); Armour-Garb and Woodbridge (2015). The six pretendisms are inequivalent.

beings like us.[47] Frictionless planes are another thing. There are adequate physical measures of the degree to which a December pre-game National Hockey League ice-surface in Toronto approximates to a frictionless plane. As long as this present omission persists, the normative salvation of the empirically false is all wind-up and no pitch. And, it hardly needs saying that there is no relief to be had in the N-N convergence thesis of a naturalized logic.

The empirically discernible facts of lived literary experience are unassimilable to settled and preferred visions of contrary and counter-evidenced providence. This has all the hallmarks of dogma which harbours the perplexing notion that a philosophical theory of an empirically discernible subject-matter is itself immune from empirical discomfort. This triggers the still unsettled question of what a philosophical theory of such orientation is answerable to. One largely unspoken answer is the normative one: Notwithstanding that everyone experiences himself as we all do in our literary engagements, we should all stop experiencing ourselves in those ways and start approximating to experiencing ourselves in the way the theory mandates. More of Priest's recommended rewiring of our understanding? What some people won't say to preserve their dim view of humanity's lowly cognitive yield!

At an earlier MBR meeting, under the title "Against fictionalism", I offered some reflections on the alliance of unrealist science with theories of fiction.[48] Part of what I said there touched on how the peculiarities of stories release disanalogies with the made-up things of unrealist science, which are more than sufficient to discourage its lodgement in the mainstream logic of fiction. My rehearsal of those literary peculiarities was somewhat sketchy and incomplete. In this sequel I would like to up the game of "Against fictionalism" and, in so doing, expose a problem about inconsistency which has yet to attract the attention that's due it.

9 Untroubling Systemic Inconsistency

Guided by the naturalistic constraints, it is transparently the case that cognitive dissonance is wholly pre-empted by the in-fiction/in-the-world dichotomy. The truths of fiction are true in the story and false in the world. It is a distinction that neither imputes nor imparts ambiguities to "true", and "false", to the sentences of its respective placements, or to any of their respective parts. Although agreeably at home with respect to truth, the distinction is no home *at all* for referring expressions. It is true that "Sherlock lived in Baker Street" is true in the story and false at home. It is true that in the story "Sherlock" refers to Sherlock, but not true in the world that it doesn't. "Sherlock" refers to Sherlock *each* time, both when Watson addresses him by name and when we remark on his deductive powers. Otherwise we in the world could not say of any same thing that he lived in Baker Street in the story and did no such thing in reality, thus missing to the uttermost peculiarity of the semantics of fiction.

[47] Frank Zenker is excellent on this point. See his "Can Bayesian models have 'normative pull'?", in Steve Oswald and Didier Maillat, editors, *Argumentation and Inference,* volumes 76 and 77 in Studies in Logic, London: College Publications, 2018.

[48] Woods (2014).

Against referentialism: This is a telling comeuppance. It puts paid to a referentialist account of truth in fiction and frees it from hostile theoretical preconceptions to the contrary.

This is as it should be. The truth-makers required for fiction are radically different from those that semanticize discourse about the world. In the first instance, Sherlock's truths are from Doyle's pen, and thereafter from all others about him which, even if not made true by Doyle, depend without exception on those that are.[49] Whereupon a second significant comeuppance. The fact that Doyle is ultimately their maker and that, when he makes them, he imposes no ambiguity upon them, whatever confers meaning on natural language sentences, also confers it on the very sentences that Doyle makes true. In making them so, Doyle gives them no meaning beyond the meaning already possessed. Unless it sets out to make it otherwise. Lines above, we saw that a referentialist semantics is unavailing for fiction. We now see that it is also so for conditional semantics.[50]

Against truth conditional semantics: The conditions that make fictional sentences true do not give them their meaning.

Stripped of theoretical rivals, the theory of truth in fiction is free to make its own empirically sensitive way. It is necessary to ask where the meanings of fictional sentences come from if not from their truth conditions? Its answer requires recognition of an important inheritance property for fiction.

World-inheritance: Save for an author's contrary provisions, stories inherit the world. Everything true in the world at the time in which the story is set is true of the world of the story.[51]

If this weren't so, literary texts would be saturated with an indeterminacy so pervasive as to guarantee their readerlessness. It also forecloses on the well-travelled abstractionist idea that Sherlock is an ontically incomplete entity, with a brother but not a mum or dad, or the occasion or wherewithal to visit the Gents at Victoria Station.[52] By those same lights, we have a further inheritance property to log.

Meaning-inheritance: Whatever confers meanings on sentences about the world, stories inherit them, save again for creative auctorial interference.

[49] For example, from "Holmes waved our strange visitor into a chair", we have it that Holmes waved someone into something, and that the something had feet and probably legs. This is London, after all, not the tent of a desert shiek.

[50] The origins of which lie in Wittgenstein's *Tractatus Logico-Philosophicus,* para. 4.024, London: Routledge & Kegan Paul, 1921/22. The passage reads: "To understand a proposition means to know what is the case if it is true." The passage is overstated. Assume that we know what would be the case if *p* were true. If *p* were indeed true, it would be the case that everything in p's deductive closure is also true. Since it has at least the cardinality of the natural numbers. Therefore we do not understand p. We cannot know what, if it were true, all that is the case, i.e., all that follows of necessity from it. Its deductive closure is transfinitely lush.

[51] This necessitates a distinction between world-sentences that are true *in* the story – e.g. that the Strand lies south of Baker Street – and world-sentences true *of* the world of the story but not *in* it – e.g. that Sevilla lay south of London in the 1880s, as it does now.

[52] Initially, the working title of *Truth in Fiction* was *Sherlock's Member: An Essay on Truth in Fiction.* Like Queen Victoria, my mother would not have been amused. I am a bit sorry now that I chickened out.

Let's briefly come back to the truth-conditional semantics remark of three bullets ago. Truth conditional semantics arises in 4.024 of Wittgenstein's *Tractatus*:

"**4.024** To understand a proposition means to know what is the case, if it is true."[53]

As now understood, the doctrine asserts that the meaning of a proposition is determined by its truth-makers, by the way the world has to be in for the proposition to be true. Lines just above, I said that because the meaning of the sentences of Doyle's making is not set by the fact that he is their truth-maker, a truth-conditional semantics is wrong for fiction. Part of the reason for saying so is the claim that stories inherit the meanings of the sentences of which their authors are their truth-makers. On the other hand, there was on that same page the claim that stories inherit the world. In each case, the defeasing qualification is understood. Except where expressly otherwise provided for by their authors, the world of the stories is the world. When we turn our minds to the question of how the world would have to have been to make it the case that Doyle (rather than Kant) is the truth-maker for *The Hound of the Baskervilles* (and not *War and Peace*), we begin to see that Doyle, the writer of his stories, is brought about by the world. What brings it about that Doyle is the truth-maker for his own fictional sentences is the way the world was in when he was exercizing his truth-making powers. From which we should conclude that Doyle's truth-making powers are parasitic on the truth-making powers of the world. In the end, it is the world that calls the shots for meaning and truth alike. It enables Doyle to call the truth- shots for fiction. It withholds its own powers to make Doyle's sentences what they mean, absent his express provisions to opposite effect. So the contra truth conditional semantics thesis requires some refinement.

The nub of fiction's most distinctive semantic problem is that every sentence of Doyle's text is concurrently true and not true, yet safe from cognitive turbulence by a distinction that imparts no ambiguity. On the face of it, this should not be so. Every untutored literate human being knows that it is impossible for something to be unambiguously true and not at once. Sentences *thus characterized* are contradictions, making it inexplicable that they occasion no cognitive bother, and also leaving some much more bothersome collateral damage. To come back to an old example, ZF set theory was purpose-built to evade the inconsistency that beggared its naïve predecessor. Had it organized itself on the model of the semantics of fiction, it would be landed in an even worse inconsistency problem. As customarily understood, naïve set theory abetted one inconsistency and, without the help of the *ex falso* theorem, none other.[54] Fiction is different. Its every sentence is unambiguously true and false from the get-go, independently of what we might be inclined to make of *ex falso.* It is beginning to look as if the semantics of fiction would be a fatefully wrong turning for unrealist

[53] Ludwig Wittgenstein, *Tractatus logico-philosophicus,* London: Routledge & Kegan Paul, 1921/1922; p. 47.

[54] According to the *ex falso quodlibet* theorem, from a logical falsehood, everything of necessity follows. I think that the theorem is provable for English, but I shan't press the point here. It is possible to show why, even if true, *ex falso* would not be a comeuppance or in the least cognitively disturbing. For details see *Truth in Fiction,* 2018; especially chapter 9, "Putting Inconsistency in its Place".

science, certainly if the customary view on the concurrency of truth and falsity held fast.[55]

If we can't wring some further clarity from the in-fiction/in-the-world contrast, we'll be confronted by accusations of floating a distinction that marks no difference. If we don't soon find some useful work for it, we'll be scorned for empty word-play. What, then, *does* the difference mooted by the distinction actually make? The answer is as plain as the nose on a properly made face, that is to say, right under our eyes. The imputed difference is not one of meanings. It is a difference of place for unambiguous sentences to be at once true and not true in, or in other words, a difference of truth-*sites.* Accordingly,

> *The truth-sites thesis*: "Sherlock lived in Baker Street" is true *in situ* the stories and concurrently and unambiguously false *in situ* the world.

Moreover,

> *What sites are*: Sites, in this usage, aren't really places. They are figures of speech in which sites are reifications of *contrary truth-makers applied concurrently to unambiguous sentences* such as "Sherlock lived in Baker Street".

The conditions making the Sherlock sentences true ensue from Doyle's scribblings in Edinburgh. The conditions making it untrue are facts made true by the state that London was actually in the 1880s.[56]

"Ah, yes", critics will say, "by your own avowals, differences *in situ* don't erase the concurrent truth and untruth of the sentences in their ambit. So, on your own telling, they are all of them contradictions of such abundance as might have tested the nerve of a Heraclitus." It is a fair point, and rightly presses for an answer. My answer can be found in what is arguably Aristotle's most influential and well-travelled idea. In the *Metaphysics,* he gives the Law of Non-Contradiction (LNC) three different and apparently pairwise inequivalent formulations, in the following order:

> *The ontic formulation*: It is impossible that the same thing belong and not belong to the same thing at the same time and *in the same respect.* (1005^b 19–20; emphasis added.)[57]
> *The doxastic formulation*: No one can believe that the same thing can (at the same time) be and not be. (1005^b 20–25).
> *The logical formulation*: The most certain of all basic principles is that contradictory propositions are not true simultaneously.

In *Truth in Fiction,* I said that the logical formulation understates the LNC. I now see that it doesn't. The logical formulation draws upon the prior ontic formulation to derive

[55] It is important to keep in mind the difference between the systemic inconsistency of the sentences of fiction and the world, and the *internal* inconsistency of stories such as the one in Ray Bradbury's "A sound of thunder", in which Keith is elected President in 2055 and not elected President in 2055. For disposal, see section 9.5 of *Truth in Fiction.*

[56] In the 1880s, the number of addresses in Baker Street was 100. No one lived at 221 then. There was no such address.

[57] Aristotle (1984).

the concept of a *contradictory sentence* - a sentence which in all the same respects is concurrently and unambiguously true and not - of which it repeats the impossibility of its instantiation, adding that this is the most certain of all basic principles. We also see that the ontic formulation vindicates the doxastic formulation, on which it is not possible to believe any proposition knowing it to be one which, in all respects, is unambiguously and concurrently true and not true.[58] "Sherlock lived in Baker Street" is different. There is one respect in which it is true, and another in which it is false. But there is no same respect in which it is true and false together. "Sherlock lived in Baker Street" is concurrently and unambiguously true and false, but it is not a contradiction. None of the LNC formulations lays a glove on it.[59]

Aristotle offers no account of respects. We might think that meanings are respects. But Aristotle's logic presupposes throughout that the meanings of its sentences are constant.[60] So I conclude that by "respects" he doesn't mean "meanings", and conjecture that he might well mean something close to what I mean by "sites". There are two clear cases in which site-qualification clauses do productive theoretical work. One is the semantics of fiction. The other is jurisprudence. When Spike is found guilty on a non-reversed verdict, he might not be guilty in fact. No matter how innocent, miscarriages of justice owe nothing to the imagined ambiguity of "Spike is guilty". The world and law are different truth-sites for such findings of fact. It is true in the law, and its negation concurrently could be true in reality. It is much less clear whether sentences such as "A rational person closes his beliefs under consequence" are amenable to truth-siteness. Are we to say that, while false *in situ* the world, they are true *in situ* classical decision theory? I own to some misgivings about saying so. Unrealists of instrumentalist leanings are likely to concede that, although that sentence is everywhere false, it helps in coming upon truths that are everywhere true (or otherwise virtuous). Yes, but: We have yet to find convincing cause to suppose that in anything their traffickers may think that they are doing with them, they are making these falsehoods *true* anywhere.

Some will think it a knock against sites that I've offered no necessary and sufficient conditions for them. Only half of that is true. I've already laid down sufficiency conditions, and that's one better than Aristotle did for respects. Besides, lamentation over the lack of necessary conditions should leave us cold. Aristotle offered a sufficient condition for the relation of following of necessity from, and went on to build humanity's first truly great logic without need of necessary ones.[61] Relevant logicians

[58] Dialethists disagree, to no avail I think. See chapter 8 of *Truth in Fiction,* section 8.5, pp. 167–168.

[59] Actually the meanings of the instantiations of its schematized terms. I might add that without a "some respects" clause the classical (model theoretic) definition of deductive consequence cannot be said to be monotonic. This is shown to be so in "The promiscuous reach of deductive consequence" forthcoming.

[60] *Categories* is the first monograph of the *Organon*. Although written in an apparently metaphysical style, it is better understood as a treatise on ambiguity, which concentrated on the several-fold ambiguity of the "is" of predication. Its opening sentence is "Being is said in many ways" [= "the 'is' of predication is multipli-ambiguous"]. The rest of the *Organon* hinges on the assumption that the predicative "is" ambiguity-free.

[61] True, he went on to lay down necessary conditions for syllogisms, but not for validity. Syllogisity, he insisted, is a special case of validity, which is wider. See *Prior Analytics* A 32 47^a 33–35.

in the manner of Anderson and Belnap offered only a sufficient condition for content-sharing relevance, and then went on to stimulate thriving programmes in relevant and paraconsistent logic.[62] Bueno gives us a usable and load-bearing existence-predicate in the cause of scientific unrealism, without the necessity of providing necessary conditions. Kit Fine's reality predicate and possible worlds idiom are used to good effect in the absence of necessary and sufficient biconditionals.[63] The causal response characterization of knowledge in Sect. 5 advances conditions sufficient but not necessary thereunto. The demand for necessarily true biconditional definitions of properties of central interest is a demand more often made than complied with. My own view of this matter is to lighten up and wait to see what happens.

At least we have a solution to the unworrisomeness of fiction's inconsistencies:

> *Solving the untroubling consistency* problem: The concurrently true and false sentences of fiction are inconsistent, *but they aren't contradictions*. They are rationally believable and logically tenable.

10 Refutations and Conjectures

One of the reasons for liking a naturalized logic for fiction is that it encompasses an epistemology that fits the data of lived literary experience. I said earlier that the epistemology I favour is the causal response (CR) variation of reliabilism which plays a large role in *Errors of Reasoning*. The parts of it that I've called upon here are the respect for data principle, the empirical sensitivity principle, the N-N convergence thesis, the Abundance theses for knowledge and error and the Enough thesis for the state of our cognitive well-being.

Taking the last three as an example, there are well-known epistemological objections according to which Enough could be true and yet Cognitive Abundance false. Relatedly is the view that while, the lived relations of readerly and writerly literary congress are experienced as facts, they needn't actually be facts. I reply as follows.

- It was Field himself who coined the catchy line that a statement or theory doesn't have to be true to be good. I accept that Enough doesn't entail Cognitive Abundance, but it does lend it decent abductive support. In a more sweeping version of Field, Stephen Stich wonder why, if Enough is actually true, "anyone but an epistemic chauvinist would be much concerned about rationality or truth[.]"[64] My answer is that this is an implicit capitulation to Big-Box Scepticism and the recurrence of the notion that when humanity reads the stories it likes, it loses its mind. These are doctrines none of whose proponents even try to oblige behaviourally after taking the bus home after a hard day at the Cognitive Nihilism Lab.
- None of the rival epistemologies fits the data of our literary engagements as well as a CR-epistemology does. The respect for data principle pledges itself to disoblige or

[62] Anderson and Belnap Jr. (1975). See also Restall (2006).

[63] Fine (2009), and Fine (2010).

[64] Stich (1990).

> scant those data only should there be weighty cause to do so. I acknowledge the commitment but not the probable cause. Any theory of fiction which scants or overrides those data hasn't a ghost of a chance of providing a coherent and compelling account of fiction.

It is an understatement to say that, in the light of Enough, Cognitive Abundance is an abductively tenable inference. Closer to the truth is that without Cognitive Abundance, Enough would be little short of a miracle. Putnam is helpful on this point. Although he was writing it in support of a realist treatment of the theoretical entities of science, it also applies here:

> *Putnam's no-miracles principle:* Philosophy is rarely well-served by espousing positions it takes a miracle to explain.[65]

Big-box scepticism about literary discourse embodies a question that draws an answer from Kant. The question is "Could it be that whenever a neurological human being engages with a story he's reading he has fallen off his cognitive perch?" The answer, albeit not directly to this one, is:

> "To know what questions may reasonably be asked is already a great and necessary proof of sagacity and insight. For if a question is absurd in itself and calls for an answer where none is required, it not only brings shame to the propounder of the question, but may betray an incautious listener into absurd answers, thus presenting, as the ancients said, the ludicrous spectacle of one milking a he-goat and the other holding a sieve underneath."[66]

The Basic Laws are stone-eyed killers of any theory of human thought or speech which strives for an empirical adequacy unruffled by the presumptuousness of the normativity poseurs. Meinongian semanticists do themselves credit by declining truck with the Basic Laws that come after Law I. But, in taking the formalizing approach, they place themselves at a greater remove than necessary or wise from the data. In their affection for formal semantics, they also expose themselves to the folly of leaving Eureka in February and wintering in Karosjok. Any theory of fiction that scants those data, or overlooks them entirely, will be significantly compromised by their omissions, if not wholly felled by them. In what remains of this paper, I'll take it as read that the naturalized logic of fiction sketched in these pages, and more fully developed in *Truth in Fiction,* is at least a *contending* theory, and is sufficiently so to serve as a test case for the modelling question for scientific theories of the unreal.

Let's call the contending theory *N-fic.* To bring this off, we can list N-fic's characteristic features, the ones that do its heavy lifting. Then we can search unrealist science – call this *UR- science* – for counterpart features playing similar roles in them.

> *Stage* one: Should the number of counterpart matchings be high, we will have scaled the first stage of our enquiry. If we don't arrive at that stage, our question is answered: N-fic is not a fruitful theory on which to model scientific theories of the unreal.

If the ratio of mismatches to matches were notably high, we would move to

[65] Putnam (1979).

[66] Kant (1965; A58/B83).

Stage two: It would befall us to lay down suitably clear criteria for what it would take to bring it about that unrealist science has a model in N-fic, or can be modelled on it.

We begin with some points of matching. N-fic and UR-science allow for reference and ascription to the nonexistent. They both allow that their purported referents are the subjects of load-bearing sentences which occur in the domains and counterdomains of consequence relations and are eligible to participate in the drawing of inferences. The UR-sciences are also at one with other semantic features of N-fic. Except for specially coined neologisms, meanings in UR-science mean what they mean in real-world contexts. They reflect an openness to world-inheritance.

Seams start to strain, perhaps to the popping point, when we come to belief and knowledge and, at their bottom, *truth*.

UR-falsehoods are believed false: It is not in general the case that the falsehoods of UR-science are actually believed by philosophical unrealists, or known by them to be true, still less believed and not believed and known to be unambiguously true and concurrently not true.

As already mentioned, "It is not really true that any rod is perfectly rigid" is not a theorem of the physics that postulates them, and "It is not really true that populations are infinitely large" is not a theorem of population genetics. On the other hand, "It is not really true that Sherlock lived in Baker Street" is an *essential* theorem of N-fic. Moreover, UR-science is not a natural home for the untroubledness of systematic inconsistencies problem, and is therefore unamenable to and unneedful of its solution in N-fic, or any simulacrum of it. From which we should conclude its failure to arrive at stage two. So let's make it official:

The model-unworthiness of N-fic for UR-science: If, as I respectfully submit, N-fic is the best logic for the discernible empirical data of our worldwide engagement with stories, it fails the counterpart-matching requirement for stage-one modelling success, and cannot therefore proceed to stage-two.

11 Lies Without Unrealism

Judging from the title of this paper, and its Cartwright epigraph, one might reasonably expect that it would judge Nancy Cartwright's pure fictions about how physics is modelled to be yet another of the follies indulged in by fictionalists of science. That would certainly have been so had Cartwright framed the theoretical models in the manner in which Tarski, for example, modelled natural language truth in a nonconservative extension of the model theory of first order logic. But, as Otávio Bueno has helped us see, for Cartwright's pure fictions to do their assigned work, it is unnecessary to hold them to a Tarski-like or referentialist semantics. Why? Because if pure fictions are modelled referentially, they collapse under the weight of the Basic Laws. Tarskian models simply don't allow for contentful reference-failure. In Sect. 1, we saw that in Bueno's version of fictionalism, fictionalist sentences are false. To this a qualification is added. Their ontological commitments, if any, would remain epistemically underdetermined. Fictionalist science would adopt an agnostic stance towards reference and referents. At first glance, the qualification might strike us as a hedge, as a way of

dodging a hard question. But this is very much the wrong reading. To see why, consider again how Cartwright characterizes the "pure fictions" of physics. In its core sense, a lie is the deliberate telling of something untrue. It need not be told with an intention to deceive.[67] Cartwright's lies, and Bueno's fictionalist falsities too, are transparently untrue, but they are advanced for a benign purpose. Let's call sentences of this sort "Cartwright sentences" and draw attention to what makes them special. They are the lies that Cartwright calls "pure fictions". It is natural to say that Cartwright sentences model things as they aren't, and in some cases could not be in nature. Modelling, here, is representing, conceiving of or a way of thinking of. When rods are modelled as perfectly rigid, the representation is false on purpose. It is often said that when those conditions obtain, Cartwright sentences are true in the model of their purported subject matters. Models here are the *reifications* of the abstract noun cognate with the activity-verb "to model." This raises what we could call "the reification of modelling question" or ROM-question for short.[68]

A Cartwright sentence in a theory $\mathcal{T}$ is a sentence to which, when taken at semantic face-value, nothing in nature approximates in any finite degree. A Cartwight theory is one whose language contains at least some working occurrence of Cartwright sentences. There are various things that Cartwright sentences aren't: they aren't assumptions; they aren't hypotheses and they aren't conjectures. They are *confirmation-abbetors,* and, save for that role, there is nothing that Cartwright's lies are on scientific trial for in $\mathcal{T}$. They perform their one and only function by way of their indispensable contributions to $\mathcal{T}$'s derived results, results on the basis of which $\mathcal{T}$ earns high confirmation ratings at the empirical checkout counter. A Cartwright sentence S_c is an untrue sentence to which full *premissory* authority is given. Although untrue, there is nothing it's in *error* about. There is nothing in $\mathcal{T}$'s subject-matter that S_c causes $\mathcal{T}$ to misunderstand or get wrong. Imagine that there were some honest scientific work to do by granting full premissory remit to "The present king of France is bald." No one at one with the semantic conventions of English (and also reasonably up to date on how France is governed) would diagnosis the fault in this sentence as implying the existence of that particular nonexistent. There is no such thing as *that particular* nonexistent. It is perfectly plain to them that there is nothing whatever to which that sentence refers and nothing of which it is false. That is what Cartwright herself takes Cartwright sentences to be. *Semantically*, they function in the way that "The present king of France is bald" has functioned since 1905. On the score of *scientific value*, however, it is open-and-shut for Cartwright sentences and a dead loss for "The present king of France". This strongly suggest a negative answer to the ROM-question.[69]

[67] Consider the baldfaced liar who lies to his interlocutor as an alternative to admitting to something untoward. The lie's secondary purpose is to shift the burden of proof to the interlocutor.

[68] Of course, there are other lexical origins of the word "model", sometimes denoting objects or structures whose value to an enterprise extends well beyond the promise of an abstract noun attaches to the modelling of something as. Think here of scale models in architecture and strolls down the runways of Paris and Milan.

[69] There is little doubt, however, that Cartwrights own answer to the ROM-question is affirmative: "My basic view is that fundamental equations do not govern objects in reality; they govern only *objects in the model*." (p. 153, emphasis added).

Consider an example. It is not actually one of Cartwright's own, but perfectly good for present purposes from population genetics: "Populations are infinitely large."[70] This, of course, is untrue. But it is needed for well-confirmed derivations about natural selection *in terra firma.* For S_c to do its job in T it is unnecessary for it to carry ontological commitments in T's domain of discourse. It is actually worse than unnecessary. If its semantic functioning required S_c's quantification over populations, then S_c's premissory entitlement would make for a biological impossibility, a trait it would pass on to T itself. In that case, T couldn't be instantiated in nature. Consider now the subject and predicate terms of S_c. Context determines that "populations" denote biological populations and "infinitely large" denotes a property of set-theoretic structures. There is nothing in T's domain that instantiates the predicate and nothing in the predicate's extension that includes the subject-term's extension. As a matter of natural and mathematical necessity combined, the two extensions have a null intersection. There is nothing whatever there.

We only have to remind ourselves that whatever it takes to take "Populations are infinitely large" at semantic face-value, it cannot be by subjecting it to a referentialist interpretation of it. If we did take that approach, we would be faced with a theory that nature cannot abide and a sentence that expresses no proposition We saw in Sect. 6 that semantic referentialism is a dead loss for the meaningful sentences of fiction. This gives us a clear shot at absorbing the good sense of Bueno's reservation about unrealist science's ontological commitments, which I take as tending towards a negative answer to the ROM-question.

Indeed, it is here that Bueno's reservation has real bite. Not only need we not speak of the ontological commitments of Cartwright sentences, we couldn't if we tried. In these cases, the question of ontological commitment cannot arise, except negatively. There is nothing in Cartwright sentences to off-load to its ontology. That, in turn, argues for upgrading Bueno's, agnosticism to a more determined ontic atheism, that is to an unqualifiedly negative answer to ROM. This would leave the question of whether Cartwright sentences are to be taken at semantic face-value, and the follow-up question of how fictionalists actually do take them. The answer to the first is determined by whether a sentence that refers to nothing and attributes nothing can qualify for the premissory latitude that Cartwrightian fictionalism grants them. On the empirical record of scientific practice, the answer is assuredly *yes*. Having been freed of late from the drag of referentialism, Cartwright sentences invite straightforward face-value semantic readings. To be clear, the face-value reading of "Populations are infinitely large", much like "The present king of France is bald" is that there is nothing whatever that it's false *of*. Of course, this is not remotely the way in which philosophers of science and language usually read Cartwright sentences. If we generalized and strengthened Bueno' reservation, we could say that the received way with Cartwright sentences is absolutely the wrong way to read them. It is, on the other hand, absolutely the right way to read

[70] More precisely, when we conceive of a population as infinitely large, we take it to be a deterministic system without stochastic noise. In nature, however, populations always have deterministic and stochastic features. This is the fact that gives the lie to the Cartwright sentence.

"Sherlock lived in Baker Street" and "Some things don't exist". There are no truth-makers for Cartwright sentences. There are two concurrently contrary but perfectly kosher truth-makers for ambiguity-free Sherlock sentences. One makes them true. The other makes their negations true. In light of these considerations, we can say with still greater emphasis that the modelling of Cartwright sentences on Sherlock sentences is the utter folly of counterproductive intercourse with alien models. But I am bound to say that was never Cartwright's intention. Cartwright's fictionalism is fictionalism in name only.

Although conventional unrealism is a wholly self-defeating response to the fruitful particularities of Cartwrightian science, this is far from saying that theoretical science is purely fictional across the board. Mathematics is not in the least Cartwrightian, but neither is it in the least unrealist. When a mathematician has an idea that interests him, he frequently looks for some useful work for it to do. Such ideas are the work-horses of mathematics. Ferreirós sees them as hypotheses, not in the speculative or conjectural sense, but rather in a sense that can rightly be called constitutive. Let H be one of these ideas. *Unlike* Cartwright sentences, they are *provisionally* given premissory free rein, and its employers wait to see how it makes things go. If H withstands all attempts to refute it and plays a growingly fruitful role in the production of new and well-approved theorems, then in time it sheds its provisional status and takes on the character of a well-received but not yet proved proposition of mathematics. Ideas this fruitful are as honey to bears, and in time become the targets of fully-fledged demonstrations. When these proofs arise, the now successfully targeted Hs takes on the gravamen of mathematical truth. Nothing like this lies in store for Cartwright sentences.

It is here that we come upon a telling similarity and difference with fiction. Like the objects of fiction, the objects of mathematics are man-made. *Unlike* the objects of fiction, there is no concurrent reality that stands against them. Correspondingly, for the man-made truths of mathematics, there is nothing antecedently or concurrently that makes them false. So seen, philosophical unrealism is the wrong accounting. There is nowhere in which the truths of mathematics aren't true, and nothing at all that they are false of.[71] Accordingly, mathematics has no fit place in the embrace of philosophical unrealism or any of its fictionalist variants. The right accounting is a realist one, a realism that grants full membership to humanly created things. *Creative realism* we could say.[72] There is more to be said of this in work underway in a project called "Making things true: A naturalized logic for mathematical knowledge."

[71] This is not to deny that mathematicians make errors or that a fair share of proofs published in reputable journals turn out to be defective. The point here is that when a mathematician's H loses its provisional status and takes on the gravamen of truth, there is no non-mathematical fact that it is guaranteed to contravene.

[72] Creative realism should not be confused with constructive realism (or construct realism) in the sense of Loevinger or Wallner. In their approaches construction or constructs are models of reality, not constituents of it. See Loevinger (1957); Wallner (1990). A better comparison is with Roberto Torretti's *Creative Understanding: Philosophical Reflections on Physics,* Chicago: University of Chicago Press, 1990. A significant part of the "Making Things Real" task will be to contrive a semantics and proof theory for Cartwright theories whose working languages blend Cartwright sentences with creatively real truths of mathematics.

Before leaving this point, let me hasten to say that this same view of realism carries over to those theoretical sciences in which the objects of mathematics have an indispensable place. However, it is not my view that just anything of a theorist's making up rises to the creative bar that mathematics does.

Acknowledgements. For stimulating discussion or correspondence and wise counsel over the years about the following matters I warmly thank: Patrick Suppes, Solomon Feferman, Alirio Rosales and Matthieu Fontaine for scientific unrealism; for fictionalism Bas van Frassen, Otávio Bueno, Alirio Rosales and Lorenzo Magnani; for Meinongianism Richard Routley, Bob Martin, Terry Parsons, Nick Griffin, Dale Jacquette, Franz Berto and Fred Kroon; for pretendism, John Searle, Dom Lopes, Fred Kroon, Jody Azzouni, Brad Armour-Garb and James Woodbridge, and Manuel García-Carpintero; for causal response epistemology and naturalized logic Jaakko Hintikka, L. J. Cohen, E. M. Barth, Maurice Finocchiaro, Lorenzo Magnani, Woosuk Park, Ahti-Veikko Pietarinen, Alirio Rosales, Bernard Linsky, Madeleine Ransom, Shahid Rahman, Juan Redmond, Matthieu Fontaine, Selene Arfini, Cristina Barés-Goméz and James Freeman; and for the N-fic semantics of fiction, Dom Lopes, Ori Simchen, Chris Mole, Dale Jacquette, Vedat Kamer, Hartley Slater, Michel-Antoine Xhignesse, Manuel García-Carpintero and Otávio Bueno. For stimulating discussions in Sevilla, I also thank Selena Arfini, Daniele Chiffi, Matthieu Fontaine, Alex Gelbert, Andrew Gibson, Lorenzo Magnani and Shahid Rahman.

References

Anderson, AR Belnap Jr ND (1975) Entailment: the logic of relevance and necessity, vol 1. Princeton University Press, Princeton, vol 2, with Michael Dunn J (ed.) 1992

Aristotle (1984) The complete works of aristotle: the revised English translation, in two volumes, Barnes J (ed). Princeton University Press, Princeton

Armour-Garb B, Woodbridge JA (2015) Pretense and pathology: fictionalism and its applications. Cambridge University Press, Cambridge

Batterman RW (2009) On the explanatory role of mathematics in empirical science. Br J Philos. Sci. 61:1–25

Bentham J (1932) The theory of fictions. In: Ogden CK (ed) Bentham's Theory of Fictions. Harcourt, Brace and Company, New York

Berto F (2013) Existence as a real property: the ontology of meinongianism. Springer, Berlin

Berto F, Plebani M (2015) Ontology and metaontology: a contemporary guide. Bloomsbury, London

Blackburn S (2005) Oxford dictionary of philosophy, 2nd ed. Oxford University Press, New York, pp 309–310. First edition 1994

Bradbury R (1952) A sound of thunder. Collier's Magazine

Bueno O (2010) Can set theory be nominalized? A fictionalist response. In: Woods J (ed) Fictions and models: new essays, with a foreword by Nancy Cartwright. Philosophia, Munich, pp 191–224

Bueno O, Colyvan M (2011) An inferential conception of the application of mathematics. Noûs 45:345–374

Burgess A (2010) Metaphysics as make-believe: confessions of a reformed fictionalist. In: Woods J (ed) Fictions and models, pp 325–343

Cartwright N (1983) How the laws of physics lie. Oxford University Press, Oxford

Daniels CB (1987a) A first order logic with no logical constants. Notre Dame J Formal Logic 28:408–423

Daniels CB (1987) 'The story says that' operator in story semantics. Studia Logica 46:73–86
Dayaratan K, McKitrick R, Krutzer D (2017) Empirically constrained climate sensitivity and the social cost of carbon. Climate Change Economics 8:1–12
Ferreirós J (2015) Mathematical knowledge and the interplay of practices. Princeton University Press, Princeton
Ferreirós J (2007) Labyrynth of thought: a history of set theory and its role in modern mathematics, 2nd edn. Birkhäuser, Basel. First published in 1999
Field H (1980) Science without numbers. Princeton University Press, Princeton
Fine K (2009) The question of ontology. In: Chalmers D, Manley D, Wasserman R (eds) Metaphysics: new essays on the foundations of ontology. Clarendon Press, Oxford, pp 157–177
Fine K (2010) "Foreword" to Guido Imaguire and Dale Jacquette, editors, Possible Worlds: Logic, Semantics and Ontology, Munich Philosophia, pp 9–14
Frigg R (2010) Fiction in science. In: Woods J (ed) Fictions and models, pp 247–287
Gabbay DM, Kanamori A, Woods J (eds) (2012) Sets and extensions in the twentieth century. In: Handbook of the history of logic, vol 6. North-Holland, Amsterdam
García-Carpintero M (2018) Review of John Woods, truth in fiction: rethinking its logic. Notre Dame Philos Rev 29:09
Gigerenzer G (1996) From tools to theories. In: Grassmann C, Gergen KJ (eds) Historical dimensions of psychological discourse. Cambridge University Press, Cambridge, pp 336–359
Goldman AI (1967) A causal theory of knowing. J Philos 64:357–372
Gray J (2008) Plato's ghost: the modernist transformation of mathematics. Princeton University Press, Princeton
Jacquette D (1996) Meinongian logic: the semantics and existence and nonexistence. Walter de Gruyter, Heidelberg and New York
Kalderon ME (2005) Introduction. In: Kalderon ME (ed) Fictionalism in metaphysics. Clarendon Press, Oxford, pp 1–13
Kanamori A (2001) The higher infinite: large cardinals in set theory from their beginnings, 2nd edn. Springer, New York. First edition in 1994
Kanamori A (2013) Mathematical knowledge: motley and complexity of proof. Ann Japan Assoc. Philos. Sci. 21:21–35
Kant I (2007) Untersuchungen über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral. GRIM Verlag, Munich. First published in 1764
Kant I (1965) Critique of pure reason, (trans: Smith NK). St. Martin's Press, New York. First published in 1787
Kripke SA (2013) Reference and existence: the John Locke lectures. Oxford University Press, New York. A reprint with emandations of six lectures given at Oxford University between 30 October and 4th December 1973
Kroon F (1992) Was meinong only pretending? Res 52:499–526
Liu C (2016) Against the new fictionalism: a hybrid view of scientific models. Int Stud Philos Sci 30:39–54
Loevinger J (1957) Objective tests as instruments of psychological theory. Psychol Rep 3:635–694
Magnani L (2018) The urgent need of a naturalized logic. In: Dodig-Crnkovic G, Schroeder J (eds) Contemporary Natural Philosophy and Philosophies, special guest-edited number of Philosophies
Meinong A (1904) The theory of objects ("Über Gegenstandstheorie"). In: Meinong A (ed) Untersuchungen zur Gezenstandstheorie und Psychologie, (trans: Levi I, Terrell DB, Chisholm, R) Realism and the Background of Phenomenology. The Free Press, Glencoe, pp 76–117, 1960

Parsons T (1980) Nonexistent objects. Yale University Press, New Haven
Pincock C (2012) A new perspective on the problem of applying mathematics. Philos Math 3:135–161
Priest G (2016) Towards non-being, 2nd edn, expanded. Oxford University Press, Oxford. First edition in 2005
Putnam H (1979) Mathematics, matter and method: philosophical papers, 2nd edn, vol 1. Cambridge University Press, Cambridge
Quine WV (1986) Roots of reference: the paul carus lectures. Open Court, La Salle, IL
Quine WV (1966) Selected logic papers. Random House, New York. Enlarged edition, Harvard University Press, Cambridge 1995
Quinton A (1957) Properties and classes. Proc Aristotelian Soc. 58:33–58
Restall G (2006) Relevant and substructural logic. In: Gabbay DM, Woods J (eds) Logic and modalities in the twentieth century. Handbook of the history of logic, vol. 7, pp 289–398
Routley R (1981) Exploring Meinong's Jungle and Beyond: an investigation of Noneism and the theory of items. Australian National University, Canberra. First published in 1980
Russell B (1967) Letter to Frege. In: van Heijenoort J (ed) From Frege to Gödel: a sourcebook of mathematical logic 1879–1931. Harvard University Press, Cambridge, pp 124–125. Written and received in 1902
Russell B (1937) Principles of mathematics, 2nd edn. Allen and Unwin, London. First published in 1903
Russell B (1905) On denoting, Mind N.S., vol 14, pp 479–493. Reprinted in Urquhart A (ed) The collected papers of Bertrand Russell, volume 4, The Foundations of Logic, 1903-05, pp 415–427. Routledge, London and New York 1994
Sainsbury RM (2010) Fiction and fiction and fictionalism, Abington. Routledge, Oxon
Searle J (1975) The logical status of fictional discourse. New Literary Hist 6:319–332
Sierpinski W, Tarski A (1930) Sur une propriété caractéristique des nombres inaccessibles. Fund Math 15:292–300
Simchen O (2017) Semantics, metamathematics, aboutness. Oxford University Press, New York
Stich S (1990) The fragmentation reason. MIT Press, Cambridge
Suppes P (1962) Models of data. In: Nagel E, Suppes P, Tarski A (eds) Logic, methodology and philosophy of science: proceedings of the 1960 international congress. Stanford University Press, Stanford, pp 252–261
Tarski A (1938) The concept of truth in formalized languages. In: Logic, Semantics, Metamathematics: Papers from 1923 to 1938, (trans: Woodger JH). Oxford University Press, Oxford. 2nd edn, Corcoran J (ed). Hackett, Indianapolis, pp 152–278
Torretti R (1990) Creative understanding: philosophical reflections on physics. The University of Chicago Press, Chicago
Vaihinger H (1924) The Philosophy of 'As If', (trans: Ogden CK). Kegan Paul, London
van Fraassen BC (1980) The scientific image. Clarendon Press, Oxford
Wallner F (1990) Eight lectures on constructive realism (= Cognitive Science I). University of Vienna, Vienna
Walton K (1990) Mimesis as make-believe. Harvard University Press, Cambridge
Woods J (1974) The logic of fiction: a philosophical sounding of deviant logic. Mouton, The Hague and Paris. 2nd edn., with a Foreword by Nicholas Griffin, vol 28. Studies in logic. College Publications, London in 2009
Woods J (2013a) Epistemology mathematicized. Informal Logic 33:292–331
Woods J (2013b) Errors of reasoning: naturalizing the logic of inference, vol 45. Studies in logic. College Publications, London. Reprinted with corrections in 2014
Woods J (2014) Against fictionalism. In: Magnani L (ed) Model-based reasoning in science and technology: theoretical and cognitive issues. Springer, Heidelberg, pp 9–42

Woods J (2018a) Truth in fiction: rethinking its logic, vol 391. In the Synthese Library. Springer, Dordrecht

Woods J (2016) Does changing the subject from A to B really enlarge our understanding of A? Logic J IGPL 24:456–480

Woods J (2015) Is legal reasoning irrational? An introduction to the epistemology of law, 2nd edn. Revised and extended, vol 2. Law and Society. College Publications, London. First published in 2015

Woods J (2018b) Pretendism in name only. Analysis 78:713–718

Woods J (2019) Four grades of ignorance and how they nourish the cognitive economy. Synthese. https://doi.org/10.1007/s11229-019-02283-w

Woods J, Rosales A (2011) Virtuous distortion in model-based science. In: Magnani L, Carnielli W, Pizzi C (eds) Model-based science and technology: abduction, logic and computational discover. Springer, Heidelberg, pp 3–30

Yablo S (2014a) The myth of the seven. In: Kalderon ME (ed) Fictionalism in metaphysics, pp 88–115

Yablo S (2014b) Aboutness. Princeton University Press, Princeton

Zalta EN (1988) Intensional logic and the metaphysics of intentionality. MIT, Cambridge

Zenker F (2018) Can Bayesian models have 'normative pull'? In: Oswald S, et al (eds) Argument and inference. Forthcoming from College Publications, London

Zermelo E (1930) Über Grenzzahlen und Mengenbereiche: neue Untersuchugen über die Grundlagen der Mengenlehre. Fundamenta Mathematicae 16:29–47



Default Soundness in the Old Approach: An Epistemic Analysis of Default Reasoning

David Gaytán(✉)

Autonomous University of Mexico City,
National Autonomous University of Mexico, Mexico City, Mexico
david.gaytan@uacm.edu.mx

Abstract. By means of an epistemological analysis of the internal components of default reasoning, and revisiting Geffner's proposal about causal asymmetries, it is given in this paper an answer to the problem of soundness in default reasoning. The kernel of that answer is an intentionalist view of certain kind of connections inside the internal structure of a default rule. This intentionalist view consists partially in assuming certain strong relationships between the prerequisite, the justifications and the conclusion in a default. Another ingredient of this perspective is the distinction between the context-dependence of a relationship and that of its *relatas*. It is offered a formal representation of the internal structure of default reasoning. Then it is built an intuitive characterization of the notion of default logical consequence. This notion is similar to the traditional notion in the sense that it attends to the challenge of clarifying the inferential mechanism of a schema, even when in the present paper the proposal differs from the traditional one.

Keywords: Default consequence · Default soundness · Causal asymmetries

1 Introduction

Reasoning about the world can be done by two basic perspectives of research: the procedural perspective and the declarative chunks perspective. In the second case, we can assume two kinds of elements: a knowledge base and an inference engine. Default reasoning is an approach capable of representing in a natural way these two elements, specially when this reasoning is represented as an interaction between rules and theoretical contexts.

Two basic representations of defaults are the conditional representation, at the level of the object language, and the rule representation, at the level of meta-language. For example, if we want to represent a default saying that typically all birds fly, we can use the form:

$$(\forall x)(Bx \rightarrow Fx)$$
$$\text{Or}$$
$$B(x) \Rightarrow F(x)$$

which could represent object language and meta-language, respectively. Because of exceptional cases, like the case of penguins or ostriches, for example, the conditional,

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 372–392, 2019.
https://doi.org/10.1007/978-3-030-32722-4_21

$\rightarrow$, and the derivability relationship, $\Rightarrow$, are usually thought as non monotonic connections: given new information, the inference could fail. We can then retract the inference. To avoid an interpretation of necessity usually associated with universal quantifiers (different to non monotonic inference), I will represent this connections even with variables, without quantifiers (following Geffner's use).

In order to get a better representation it is important to add exception conditions as abnormal conditions in both schemas. If x could be a penguin, and we represent this exception by "ab1", then we could write:

$$\mathrm{B}(x) \wedge \neg \mathrm{ab1} \rightarrow \mathrm{F}(x)$$
$$\text{Or}$$
$$\mathrm{B}(x) \wedge \neg \mathrm{ab1} \Rightarrow \mathrm{F}(x)$$

And with this notation we want to say that, except x be a penguin, if x is a bird then x flies. This kind of rule is thought of as interacting with a theoretical context, originally configured in terms of a default theory, like in [1]. <D, W>, such that D is a set of defaults and W is a set of base knowledge about the world.

Nevertheless, there are a lot of problems we can find in the development of such a representation. In the case of the Reiter proposal, we have, in the first place, lots of problems with the proof-theoretical perspective. We have, in this respect, problems in computing the content of extensions of the original default theory. Other known problems are the interaction among defaults and the capability that an artificial intelligent agent could have to choose one among these rules, namely: the default choice problem. Another interesting problem related with the choice problem, is the problem about how to determinate the soundness of a default rule and what is the meaning of "soundness" in the case of defaults. Finally, if we are interested in the notion of logicality, we ought to lead our reflections to the notion of soundness in defaults, in order to clarify the notion of consequence supposed by them.

I think that an answer to the problem of soundness in defaults could be a way to solve the problem of the default choice problem. In this work I will deal neither with the computing problems nor with proof-theory problems. In this work I will present a philosophical analysis of the internal structure of a default rule and, using this analysis I will defend the following thesis: A rule perspective of representation of defaults in artificial intelligence, with a particular additional intensional approach, could help us in building an adequate theory of soundness in default reasoning. To clarify the strategy: One tends to think about the rule schemes in a system, *prima facie*, as good argumentative schemes; but in the case of defaults, what appears to a good scheme is not very clear at all. Intuitively, this is just the question about the soundness of the notion of derivability by default. If we reach answers about this problem we will be in a better situation to select adequate defaults for a particular decision task. Of course, I hope that this will locate us in a better situation to solve the interaction defaults problem, even when I will not treat this problem in this paper.

Normally, the problem of choosing defaults takes the form of an analysis of the interaction among defaults, trying to build some form of priority among the schemes representing defaults. An early example is [2] and a later one is [3]. Although in the last case the analysis is deeper and involves additions to the language and changes in the

internal structure of a default represented as a rule, the kernel of the analysis is what we could call, "an external analysis", since it depends on default's interaction. Héctor Geffner built an epistemic explanation of rules in order to solve the problem of default's soundness in Artificial Intelligence (AI). He saw very clearly that AI approaches face the mathematical difficulties, and the non-monotonicity of defaults but not, directly, the epistemological problems[1]. And he saw the discrimination problem as a means to solve the soundness problem[2]. To illustrate the discrimination problem of defaults I will now use Geffner's characterization. Let's see two central paradigmatic cases in this last proposal. I will use the cases, but I will explain them in a different manner in order to emphasize the problems associated with causality and the discrimination problem.

We can consider the following examples of problematic interactions between defaults:

(Shoot)[3]

Ds1) loaded $\rightarrow$ loaded'
Ds2) alive' $\rightarrow$ alive''
Ds3) shoot' $\wedge$ loaded' $\Rightarrow \neg$alive''

(Car)

Dc1) turn_key $\rightarrow$ starts
Dc2) turn_key $\wedge$ battery_dead $\rightarrow \neg$starts
Dc3) lights_were_on $\rightarrow$ battery_dead

In the Shooting case it is supposed the use an accent for the properties in order to distinguish different unities of time in the actions. We can consider that Ds1 represents the conditional "if the gun is loaded at time 0, then also is loaded at time 1". Similarly, Ds2 says "If the person S is alive at time 1, then she is alive at time 2". Finally, Ds3 says "If the gun is loaded and shot at time 1 then the person S is not alive at time 2". Ds1–Ds3 are intuitively well founded defaults. Nevertheless, if we were to have the antecedents of the conditionals Ds1 and Ds2 and also the requisite of the rule Ds3, those facts would lead us to a contradiction: the person S is alive and not alive at time 2. There are no priorities in the Shooting case, therefore we need make a decision about the use of the different defaults Ds1–Ds3. If shooting rule (Ds3) is rewritten in an equivalent way, the problem persists. Ds3 can be rewritten, equivalently, as follows:

(Ds3 Eq)

shoot' $\wedge$ alive'' $\Rightarrow \neg$loaded'

In this case the contradiction is: the gun is loaded and not loaded in the time 1.

[1] Geffner, p. 3.

[2] Idem.

[3] This is a very well-known case, called the "Yale shooting problem".

It is important to observe that there is a logical symmetry between Ds3 and Ds3Eq, and that this symmetry corresponds to a **causal asymmetry**. Geffner says that[4]:

> […] the conjunction of shoot' and loaded' *explains* ¬alive", yet the conjunction of shoot' and alive"[5] is evidence for but does not explain ¬loaded'.

In the Car case we can intuitively establish priority of Dc2 over Dc1, but not between Dc1 and Dc3. Even in the case we would prefer the worlds with Dc2 over the worlds with Dc1, we can observe anyway that, if we assume turn_key, from a conditional point of view, we will have the disjunction ¬starts ∨ ¬battery_dead, as a conclusion, "but not the stronger, *expected*[6] conclusion ¬starts"[7]. It is important to note, in addition to Geffner's reflections, that in this example the different relationships associated with the causal asymmetry could be necessity (and not only evidence in the weak sense) *versus* explanation.

Causal asymmetries have been much discussed in the Philosophy of Science, but not so much in Artificial Intelligence[8].

With these examples, Geffner's idea is to construct an epistemic explanation of rules for the interacting defaults in Artificial Intelligence. His proposal is a very deep analysis, epistemologically speaking, about the problem in this field of study respect to his predecessors. His intention is to clarify the criteria for soundness in default inference. He makes a very good analysis, at least, respect to the proof-theoretical approach, considering causal relationships and probability. The analysis presented here, even if limited to a causal approach, is similar to Geffner's analysis. But, in spite of this, the perspective I want to point to is radically different from his perspective.

In short, my proposal is that a way out of the problem of causal asymmetries can also be considered as a contribution to solve the problem of soundness in defaults. And this contribution could be built in a similar way it is usually solved the corresponding problem of deductive argument schema (or argumentation). From now on, I will call this way to characterize the logical consequence "the old approach". Let's think about it.

In the first place, it should be noted that there are at least two ways of approaching this problem: either we take the interaction between several default schemes as the unit of analysis (here the unit of analysis is the default argument, in the sense of an argumentation that uses several defaults, which implies interaction between several schemes or formulas); or we take the default schema as the unit of analysis. (In this case, the unit of analysis is the default argument, in the sense of the assumed derivability relationship inside a default structure, be it in a conditional form or in a rule form). In the first case, argumentation, one can account for a certain idea of consequence, centered on the idea of a set of consequences (that is, centered in what kind of

[4] Geffner, p. 95. Italics in the original. The distinction between to be evidence and to explain is attributed to Pearl by Geffner. The paper is: Pearl, J., Embracing causality in default reasoning. Artificial Intelligence, 35, 1988.

[5] There is a little misprint in the text of Geffner here: what appears as ¬alive" in the original ought to be alive" (without negation). I am writing it without negation.

[6] The emphasis is mine.

[7] Geffner, p. 96.

[8] For this discussion in Philosophy of Science you can see classical texts as [4] and [5].

things can be inferred). I think of this perspective as an extensionalist perspective. In the second case, focused on the idea of a logical consequence, we take the default schema as an immediate consequence (we are focused in what kind of relationship must be fulfilled by the primitive schema). I think of this perspective as an intensionalist perspective. The first kind of research provides us with the idea of what is a good inference in a default theory; the second one tells us which default is a good default or a correct default. Normally, one should be able to approach a research about a logical consequence from both points of view. This is like an assumption of correspondence between an extensional definition of logical consequence and an intensional one. It can be argued that what one wants to know in an epistemic analysis of logical systems is not only what is inferred but how it is inferred. Epistemologically speaking that difference is very important.

An investigation into the problem of the soundness of defaults, directed only in the first sense (the extensional) will not necessarily provide us with criteria of soundness of a default schema. The inverse is more plausible: providing a criterion to identify sound defaults and, this way, a way to choose among defaults organizing the interaction among them.

2 Reasons to Support a Rule-Representation of Defaults

Let's think more carefully about this. Some philosophers have divided the characterization of the logical notion of consequence as a search about the intuitions underlying that notion. In the case of [6], we can note three steps in this process of clarification: intuitions, pretheoretical approach, and explanation. A clarification of the notion of classical logical consequence, which is deductive, is usually built on the structure of these three items:

> (I) Intuitions. The notion of logical consequence consists of two ingredients: In a sound argument the inference is necessary and, In a sound argument the justification of the inference is given by its logical form.
> (II) Pre-theoretical Definition. A sound argument is one with this property: if premises are true, necessarily the conclusion is true. And this necessity is given by virtue of the connection between the logical form of the premises and the logical form of the conclusion.
> (III) Explanation (informally described). Within the framework of a formal system in which axioms can be interpreted as true in virtue of their form, and whose rules can be interpreted as preserving truth by their form, something is a sound argument if there is a formal proof of its conclusion from the set of its premises (Syntactic Approach). An argument is a sound argument if and only if all Tarskian models of its premises make its conclusion true (Semantic Approach).

I–III items can be applied to both a logical deductive argumentation and a particular deductive argument. I–III are part of an intuitive characterization of logical consequence. They are not technically semantic or technically syntactic proposals to understanding the notion of logical consequence, but underlying intuitions to these technical approaches to the notion. That is what I am calling a characterization of the

"logical consequence in the old approach". In this paper I will deal only with its syntactic approach. In this sense a rule can be analyzed as a sound deductive argumentation as much as a logically valid schema. But, in the previous characterization III, rules are assumed as pre-conditions to syntactic classical explanation. I think that the soundness problem, in an argument approach, could be viewed as a problem of the justification of a rule. In this sense, the kernel of the investigation comes out in the question of what kind of relationship is maintained in the internal mechanism of a rule. In the case of default rules, there are instances in which the connections between the set of premises and the conclusion do not correspond to necessity based on logical form: its mechanism to do the inference is based neither on intuitions (I) nor on pretheoretical definition (II). What kind of validity or correction about defaults could be characterized in the old approach?

I will reserve the term "validity" to define the particular relationship between premises and conclusion so that it satisfies intuitive items I and II. For the cases that do not satisfy those items we can talk about "correction" or "**soundness**", and we can consider validity as a particular case of correction. In non-deductive arguments soundness is usually characterized as context-dependent. Indeed, sometimes this property has been thought of a peculiar note of non-deductive arguments.

The first reason to propose a rule representation of defaults is that II and III characterize a property at the meta-logical level: the existence of a proof or the satisfiability. A representation of the defaults in the object language does not capture the second level of an argument. Given the necessity intuition (contained in I), in the standard explanation of argumentation, III does not require just a proof but a **formal** proof. Typically, the proof is expected to be constituted by rules. And these rules, in the logical tradition, provide a mechanism to guarantee truth preservation. In the case of defaults, then, we have two requirements: representing defaults in a second level of language and characterizing the transmission of truth in defaults.

The second reason to choose a representation of a default as a rule is related to an epistemic perspective of the logical consequence, in particular, related to the preference to represent the justification relation inside an argument schema, instead of an extensional perspective of that notion.

As it is known, the kernel of a characterization of the notion of logical consequence is a set of the two conditions described in item I: necessity and formality. To build an alternative characterization along the lines of these traditional conditions, and in the frame of an epistemological justification of the inference mechanism, we need a clarification of the internal elements of the default rule schema and of their interaction with a context. The formality condition appeals to an internal structure in the rules that guarantee truth preservation. My confidence rests on the fact that we can look for a similar internal condition of **formality**, into default schemas, such that we could find some similar relationship to that of **truth preservation**.

Although the context is a very rich entity and may be partially considered non propositional, an approach to some primitive kind of context could be useful in my analysis. A basic and primitive way to characterize context is to take it as a set of propositions about the state of the world, what is known as a knowledge base constituted by propositions.

In this line of thought, the formal structure of a default, inserted in a theoretical structure playing a role of a context (the default theory <D, W> in Reiter's terms [1]), it appears as a better way to begin an analysis of the soundness of non deductive arguments, which are associated to contexts. In particular, to represent default reasoning by means of rules (like Reiter's rules) appears a better way to analyze its logical consequence.

Up to this point I have centrally argued in favor of a representation of a default as a rule, as an instrument to explain the soundness of the default reasoning.

3 The Internal Analysis of Defaults and an Additional Reason in Favor of a Representation of Default Reasoning as a Rule

In order to do this analysis, it is important in first place, to note the opacity of the relationships between the components inside a default, such as it was presented in Reiter's seminal work. I will develop this section structuring the analysis on each component in relation with others components of a default. With a lot of changes and additions, and focused on other problems, the analysis and examples I will present here were developed in my [7] which was a first base to the problems discussed in this paper.

3.1 A Default Rule

To model the jump to the conclusion does not offer any problem from the standpoint of a non deductive derivation, but the interesting thing is to do the jumping procedure in a justified way. A default rule should represent a minimal justification for this procedure. We can call this procedure "the inferential mechanism".

In Reiter a default rule is presented as:

$$\frac{\alpha : M\beta_1 \ldots M\beta_n}{\omega}$$

Where α is considered as a prerequisite, each instance of β_i in the sequence $\beta_1 \ldots \beta_n$ is called a "justification", and ω is the conclusion we obtain by default, which we could call "**conjecture**". **M** is a meta-language operator which, applied to a formula, means that the formula is consistent with a certain set of propositions.

As it is written in [1], let Γ be a function on a set of formulae S relative to a closed default theory $\Delta = <D, W>$, such that Γ result in the smaller set $\Gamma(S)$ satisfying the next conditions D1–D3:

(D1) $W \subseteq \Gamma(S)$
(D2) $Th_L(\Gamma(S)) = \Gamma(S)$
(D3) IF $(\alpha : \mathbf{M}\beta_1 \ldots \mathbf{M}\beta_n/\omega) \in D, \alpha \in \Gamma(S)$, AND $\neg\beta_1, \ldots, \neg\beta_n \notin S$, THEN $\omega \in \Gamma(S)$ [9]

[9] [1] p. 89.

Γ characterizes the default consequence notion in Reiter's logic. Th_L is the tarskian function of theoremicity associated to first order classical logic, L.

We can think about this mechanism in a clearer intuitive view. A default rule would consist of a 3-phase procedure. [authorization]-[applicability or check of adequacy]-[application]. Obviously, from a logical point of view, this perspective is just metaphorical but thinking this way could help us to understand the justification process in a default. This perspective clarifies the D3 condition of the consequence restrictions in the original text of Reiter. Another consequence in this line of thought is a reinterpretation of the idea of the justifications in a default. Usually, the modality is interpreted as a clause to establish that "it is consistent to assume such and such formulae". But the meaning of this characterization is not very clear. Frequently its interpretation is that we can infer the conclusion whereas we have no evidence against that inference. But what does it mean? The minimal meaning of this is that the justification is consistent, but which set should be considered in order to check this consistency? The conclusion takes part in that checking of the consistency only in the role of one additional element in the total set of consequences. In defaults the consistency is thought to be verified between the justifications and "final" extensions. To completely understand whether it is a "final" **extension** or simply, an extension, we need to remember that it is a fixed point in the consequence function defined before. A way to describe the construction of an extension appears in the following theorem of Reiter's system[10].

Theorem 2.1. *Let* $E \subseteq L$ *be a set of closed well formed formulae, and let* $\Delta = <D, W>$ *be a closed default theory. We define:*

$$E_0 = W$$

And for every $i \geq 0$

$$E_{i+1} = Th_L(E_i) \cup \{\omega | \alpha : \mathbf{M}\beta_1 \ldots \mathbf{M}\beta_n / \omega \in D \text{ where } \alpha \in E_i \text{ and } \neg\beta_1, \ldots, \neg\beta_n \notin E\}.$$

Then E is an extension for Δ *iff*

$$E = \cup_{i=0}^{\infty} E_i$$

This is the way to represent the fixed point: an extension in Reiter's system.
In that manner, a default theory can be extended in alternative sets of consequences. This characteristic makes Reiter's system a peculiar system of logic. Partially, the difference compared with other systems is a debt to the rule perspective of Reiter's system. Reiter writes an example to show a consequence of the difference between his system and McDermott and Doyle's logic. The example is the comparison of a default theory with a non-monotonic modal theory. Although the initial elements of both theories are analogous, their fixed points do not coincide. The example assumes the

[10] *Idem.*

notion of normal default: A default in which the justification is the same (without modality) as the conclusion. The example is the following: [11].

Default Theory $\Delta = <\{A : \mathbf{M}B \,/\, B,\ \ C : \mathbf{M}D/D\},\ \{A \vee C\} >$
In this case, **Δ** *has only the extension*: Th($\{A \vee C\}$)

Non-monotonic theory $\mathbf{D}^* = \{(A\&\mathbf{M}B) \supset B, (C\&\mathbf{M}D) \supset D,\ \ A \vee C\}$
In this case, **D*** *has only the fix point*: Th($\{B \vee D\} \cup D^*$)

The first conjunct of each conditional in D* is given neither as initial data nor as a result of any derivation. Similarly, we do not have the prerequisites for the defaults of **Δ**. There is no constructive dilemma rule that allows us to infer the disjunction of the consequents of the defaults. In contrast, there is such a rule for the case of the conditionals of the non-monotonic modal logics. For both theories, **Δ** and D*, all the rules of classical logic are assumed.

The interesting thing about the example that Reiter uses to indicate the difference between both types of non-monotonic logics is that it is directly related to a way of looking at the problem of inferring from insufficient information. I think this problem, the problem to make a conjecture, to jump to conclusion, at least with epistemological skills, has been generally evaded in papers about non-monotonicity, because the authors have concentrated on the problem of how to retract inferences (the problem of retractability representation).

The kernel in an attempt to elucidate the inference in a default schema is just an attempt to clarify the mechanism of a conjecture. Systems of logic need a hard epistemological scrutiny in order for their underlying intuitions to make sense and then to improve our knowledge of the consequence relations involved.

Although the difference between both logics can be based on the use of principles or rules of the constructive dilemma type, this cannot be taken as the fundamental reason for the difference. It would suffice to add an analogous meta-principle for the defaults of a theory, to equate the results of its fixed points. One might think that it would be rational to introduce such a principle for defaults. Reiter does not choose this solution and only emphasizes the difference based on that example.

3.2 From Prerequisite to Justifications

Let us now think about the relationship between the prerequisite and the justifications. Both seem to be independently presented in the logic of reasoning by default. But the difference that Reiter emphasizes between non-monotonic modal logic and default logic could indicate something contrary.

In D* the "justifications" are part of the object language. This allows the derivation to be performed in combination with the A v C disjunction. On the other hand, in Δ the situation is very different. Even though the justifications have been met, the A v C disjunction cannot match with the prerequisite. My hypothesis is that this is not the real problem. The problem does not consist in lacking a meta-principle or a meta-rule of a

[11] [1] p. 93. The example is put in the body of the text, in the penultimate paragraph on page 93, after the proof of the theorem 2.6.

kind of constructive dilemma that is applicable to the default rules because, even having it, it would not be sufficient for the corresponding justifications to be fulfilled.

Recall the metaphor [**authorization**] - [**applicability**] - [**blocking or application**]. This metaphor also underlies the preliminary explanations of Reiter's text. In addition, the metaphor is consistent with a logically more precise intuition: the idea that the set with respect to which the prerequisite must be taken is not necessarily the final extension and almost never is. Intuition appears more clearly in Theorem 2.1 above of Reiter's logic.

In that theorem, the extension is described as the union to infinity of a series of sets of consequences $E1, E2, \ldots E_{\infty}$ that have definite relationships with each other, and such that E_{∞} is the fixed point of the function Γ. In this series, for a set of consequences E_n the prerequisite must be fulfilled with respect to the set E_{n-1}. This can be seen in the definition of the set of consequences E_{i+1}. Such a set of consequences is constructed with all the propositions that the previous set of consequences in the sequence is capable of producing, namely, E_i, plus what results from the defaults whose prerequisite is in this same set of consequences (E_i). It is interesting to note that this does not happen with the checking of justifications of the default. Its consistency is checked with the final extension, E. Thus, although from the point of view of the derivation the prerequisite and the justifications are both equally necessary to apply the default, the logical description of Reiter seems to suppose the epistemological precedence of the first with respect to the second for this application.

Of course, for a default theory, the construction of an extension starting from the verification of the prerequisites and then checking the consistency of the justifications, can be considered extensionally equivalent to a construction that proceeds first by checking the justifications and then by the confirmation of the prerequisites. However, psychologically speaking, in deciding the application or not of a default in D of Δ, an epistemological precedence of the prerequisite with respect to the justification it seems to be presupposed. This precedence to which I refer is analogous to the precedence of the requisites of a deductive rule with respect to the consequence obtained by that rule: Assuming there is no other way to get the consequence of a rule than to follow the rule, the confirmation of the requisites precedes this consequence.

The name that Reiter grants to the letter α of the general scheme of a default, namely "prerequisite", betrays this psychological burden that is, logically, unnecessary. It places the justifications as a requisite of a rule, which depends on the confirmation of another requisite: the prerequisite.

However, perhaps the idea of a rule that contains two levels of requisites, prerequisite and requisite, may be useful for modeling purposes. The idea of the epistemological precedence of the prerequisite with respect to the derivation, suggests that the prerequisite authorizes (or "pre-authorizes") the application of the default. And under this idea of authorization lies the assumption of a close connection between prerequisite and justifications. In this line of thought, once the default has been declared authorized, then it is necessary to check applicability: to check whether its exception is met or not and then to verify whether it is applied or blocked. In this way, the legitimacy of the fulfillment of a derivation would be conditioned by (a) the previous authorization and (b) the checking adequacy to trigger the default consequence (consistency checking).

In the case of the difference between D* and Δ we would have, under this interpretation, the following: There would be a strong dependence of each phase of the triggering process with the fulfillment of its corresponding prerequisite in each default in Δ and the compromise of checking adequacy (by consistency in this case) of inference. And since the disjunction A v C does not assume in particular any of its disjuncts, it would not be possible to legitimize the step to the next phase: the fulfillment of the justifications, even in the case that the disjunction were fulfilled. To say it quicker: If the prerequisite has not been given, there is no authorization for the applicability of the default. Therefore, it cannot be applied since it does not even have the property of being applicable. On the other hand, the retractable conditionals of modal non-monotonic logics lack this connection of dependence between antecedents and justifications but, even more, they lack this 3-phase triggering process of inference. In conditionals the relationship supposed in the first two steps is, simply, a conjunction. And this is the announced additional reason to keep a rule perspective in an investigation of the default consequence.

Regardless of whether Reiter maintained this assumption or not, which we would know after a difficult and obscure hermeneutical research, I will use this possibility as a pretext to find a new way of understanding the problem of conjecture (the problem of finding a sensible mechanism to jump to the conclusion).

In the case of the internal structure of a default, the proposal I want to show in this paper is a new way of understanding the interaction between the prerequisite and the justifications, but also between the justifications and the conjecture.

3.3 From Justifications to Conjecture

There is nothing in the general formal structure of a default that indicates a strong connection between the prerequisite and the proviso, nor between the justifications and the conjecture. In both cases, in opposition to the deductive tradition, formality is disconnected from the "force" of the inference. However, all the concrete examples of Reiter's text seem to imply such a connection[12]. Under all its examples lies the intuition that it is sensible or rational to think of the proviso given the prerequisite, in a certain default. In contrast, let's see the following example (a normal default case):

Ma: "My neighbor is Mexican"
Ya: "My neighbor is from Yucatán"

$$\frac{\text{Ma} : \mathbf{M}\text{Ya}}{\text{Ya}} \qquad (Y)$$

Actually, there is no reason to suppose **Ya** from **Ma**. A default like (*Y*) is a conjectural scheme much less sensible or rational than defaults presented in Reiter's text. But the difference lies not under its logical form but under its correspondence with expectations we have about the events referred by the formulae in the reasoning.

[12] [1] pp. 81–86.

We can keep a stronger connection between both prerequisite and justifications, but not between justifications and conjecture. Think about the next example.

Ma: "My neighbor is Mexican"
Ta: "My neighbor likes tacos"
Ya: "My neighbor is from Yucatán"

(T)

$$\frac{\text{Ma} : \mathbf{M}\text{Ta}}{\text{Ya}}$$

In case (T), our expectation of a connection from **Ma** to **Ta** is more plausible. Nevertheless, we only have very poor expectations "traveling" from **Ta** to **Ya**.

3.4 From Prerequisite to Conjecture

Let us now think about the relationship between prerequisite and conjecture. With this it will be easy to see also a connection between the proviso and the conjecture. The prerequisite is a necessary condition for the application of the default. If we have not reasons against that application, the prerequisite leads us to the conclusion. In addition to this relationship of derivability, at least in strictly logical terms, there is no relationship between the prerequisite and the conjecture. There is no constraint in the general scheme of a default, which indicates that the prerequisite must have some relation to the conclusion. The scheme (Y) is also an example for this kind of case: There is no connection between what the prerequisite says and what the conclusion says at least from the point of view of our expectations. Let's see an extreme example (a non normal default case):

Ma: "My neighbor is Mexican"
Pp: "Pepe is a penguin"
Dm: "The world stock market is destabilizing"

(P)

$$\frac{\text{Ma} : \mathbf{M}\text{Pp}}{\text{Dm}}$$

Definitively, this example would appear to be a case foreign to common sense. Note, however, that for each of these defaults we can imagine contexts in which someone would expect (strange as it may seem) the conclusion from the premises, based on the prerequisite and the proviso. In this sense, the logic of reasoning by default models correctly the reasoning of common sense although it does not restrict the structure of the inferential mechanism by formal connection relations. It captures common sense rationality as contextual conjectural reasoning in very general and abstract terms.

The interesting point is that, as can be supposed, Reiter does not use this kind of strange examples. As I have anticipated, the reasoning that ought interests to us, and that is presupposed in non-monotonic logic systems, is the sensible rational conjectural reasoning. Thus, in a default the prerequisite should be in a strong relationship with the justifications. The prerequisite should legitimize not only the conclusion but also the justifications. Similarly, the justifications also should be in a stronger relationship with the conclusion or conjecture and, generally, through such a relationship, the relationship between the prerequisite and the conjecture should be established.

Under the perspective of a stronger connection between the elements of a conjectural reasoning, the way of understanding the justifications changes. In the case of defaults reinterpreted from this new perspective, the caveat of the justifications not only ought to fulfill the function of determining the application or blocking of reasoning but also that of epistemologically legitimate the default conclusion.

Thus, although the prerequisite and the justifications have stronger relationships, if the justifications do not maintain any connection with the conjecture, the default would still appear disjointed.

The new way of approaching the problem of conjecture is a necessity for those who want to model more interestingly some specific kind of conjectural reasoning. No doubt, the formal instruments built so far can serve as a first step for a task like this. However, the logicians of non-monotonicity have paid little attention to stronger relationships, and the question of elucidating and putting into formal terms this kind of relationship, even if it had been only in a general form, has not been tenaciously addressed by them.

4 An Epistemic Proposal to Understand the Internal Mechanism of a Conjecture

The counter-examples presented by Geffner are effects of this essential opacity of the defaults. The reason of that opacity is the disconnection among the internal components of a default, revised in the last paragraphs. That disconnection is a formal disconnection. We normally make reasonings in common sense assuming certain stronger relationships among those components, but that relationships don't always have formal character. To reason without regard to them is not a good epistemic strategy to make a conjecture, that is, to jump to the conclusion. But, looking for a solution to this representation problem in an extensional approach means also that we do not recognize common sense assumptions.

This implies an examination of the virtues of the mechanism of inference of these schemes: an examination from the relationships between their components. This is why it is so important to build a representation of defaults that is not opaque.

In the case of Geffner, as we have seen, the problem of soundness is analyzed from the perspective of argumentation, or focused on argumentation. In spite of this, Geffner argued in favor of one kind of default mechanism of interaction among defaults, appealing to different kinds of internal relations in a default. Two key assumptions in his analysis are their distinction between explanation and evidence, and their identity between explanation and causality. However, Geffner does not intend to contribute

formally to a decrease in the internal opacity of a default. On the other hand, the assumptions from which he start can lead us to an epistemological analysis about defaults that is wrong and, therefore, misguided to solve the soundness problem from a traditional point of view (on the old approach). It will be necessary to turn towards a more decidedly intensionalist approach, to account for the inferential mechanism assumed in a modeling of default reasoning.

Geffner's basic intuition is correct: a more relevant relationship that connects our reasons with our conclusions in an argumentation based on defaults is convenient. However, the notion of evidence, as opposed to that of explanation, is an epistemologically problematic way to do that. There is, in Geffner's attempt, an aim to distinguish defaults, in terms of inference and causation. But the notion of evidence could be also theoretically connected to causation. That is, we can plausibly infer that we will have an effect and in this sense some would claim that such an effect is evidence of the existence of a particular cause. However, we could also postulate a relation of evidence based on the fact that we have a causal factor: the appearance of a causal factor can be evidence in favor that the effect plausibly had or will take place.

The problem in the analysis is produced by starting from a notion of necessary and sufficient causation, and from a notion of evidence that is necessarily assumed as not causal or, at least, if it is thought causally, it is necessarily placed as in a reasoning towards behind. Additionally, the explanation cannot be identified with causality. Although it is true that examples of causal asymmetry abound in the literature of philosophy of science, much of the discussion about scientific explanation also assumes that there could be many other relations of explanatory relevance, which would not be causal.

One escape route toward dissolving these epistemological problems could be to consider that there can be reasoning, let´s say, direct or inverse, with causal relationship; and also, direct or inverse relationship among the statements in the explanation. In both cases it can be "causal" (towards the cause) or let's say, "effectual" (towards the effect). Another difficulty is that any of these arguments could be thought of as a movement "towards the explanans", or as a movement "towards the explanandum". So the complexity is greater than it seemed.

Understanding the problem of conjecture focusing on internal components also reflects a different way of understanding the problem of soundness. The way of determining the factors that hold relevant relationships in a given situation can be addressed in an operational sense or in an inquisitive sense. In the first sense, it is enough to build an inferential mechanism that allows the jump to the conclusion, as for example the original mechanism of a default. In the second sense, we are interested in the factors that would be taken into account in the inference of a conjecture. The first sense is interested more predominantly in the result of the inference. The second one is interested in the epistemological justification of that inference. I am interested more predominantly in the kind of justification used than in the result of the inference.

The wisdom of conjectural reasoning can be sought in the epistemological relationships that may be required among elements of it. A way to face the problem, I propose, is to elucidate connections that are not entirely logical among the elements of conjectural reasoning.

One first alternative is to think of the internal connections of a default as lexical connections[13]. An example is the kind of connection between to be killed and to be dead. In a particular case, (i) one could infer to be dead based on to be killed. Or (ii) one could infer to be killed based on to be dead. In the first case, (i), the inference is necessary. In the second case, (ii), the inference is not necessary. The cases of default inference that are not susceptible to be put in deductive terms are like the second case (ii). Is in this kind of cases in which it is important the meaning of conjectural inference. Nevertheless, to assume lexical connections is neither sufficient to integrate the role of context in non-deductive reasoning, nor sufficient to clarify the internal mechanism of the inference. We would need to clarify the variations implied by context changes.

The approach I am proposing is focused on capturing connections of the referents of propositions involved in a reasoning, especially in the case of conjectures. I will call these apparent propositional relationships, "**propositional connections**". I understand them as relationships among the events referred to by the formulas of the reasoning or among, on the one hand, the events that would make some of those formulas true, and on the other, the events that would make other of these formulas true. In simple terms, the propositional connections would be relations among events described by propositions and not among propositions themselves. I think that if we focus on the reference of the propositions components of an inference, we could build a better way of control about the contextual dependence in the reasoning.

Of course, the three approaches, namely, focusing on logical relationships among propositions, that of Reiter, and focusing on lexical connections, and my own approach focused on propositional connections, are compatible. But the two last ways of thinking about internal mechanism of a default implies a greater degree of difficulty. This is because of the determination of the connections assumed among semantic links (lexical connections) or, in my own proposal, among events: Both kinds of connections are generally a context-dependent matter. Also, both suppose non logical connections among the components in default reasoning.

5 Default Soundness in the Old Approach

Let's return now to the I-III way to look for a logical consequence characterization. I will present an approach to understand default soundness attending the internal components of a default rule schema and trying to specify intuitions similar to that of the deductive logical consequence. The kernel of this attempt will consist in assuming that a way to solve the problem of conjecture is to clarify the opacity of defaults by means of propositional connections of the kind we presented above, and not only by logical propositional relationships.

The notion of default logical consequence partially consists of a formal component. Nevertheless, the connection between the elements of a default rule is not based only

[13] About this alternative, and this example, I am in debt with an anonymous referee of my paper. I had not thought this possibility before her/his direct suggestion.

on its formality. The notion also appeals to propositional connections according to a particular contextual point of view. These connections are ontological constraints. These constraints have a contextual dependence that, I propose, is directly related to the referents of the propositions involved in the default. If we are capable to represent into a default reasoning (as a default rule) this kind of connections in a formal way, we can keep a general formal structure of the default. There is at least one way to keep this formal structure without lost of contextuality, that is, there is a way to formally represent default reasoning that keeps both a formal component and an ontological and contextual component. This is the kernel of the proposal to define default logical consequence in the old approach. This proposal will allow us to define, according to I-III, (i.e., intuitions, pre-theoretic definition and classical explanation) a default logical consequence.

5.1 Intuitions of Default Logical Consequence

Let's think about (Shoot) counter-example of Geffner.

(Shoot)
Ds1) loaded $\rightarrow$ loaded'
Ds2) alive' $\rightarrow$ alive''
Ds3) (rewritten in Reiter´s style as):
shoot' : $\underline{\mathbf{M}(}$loaded') $\Rightarrow \neg$alive''

The last default rule should be equivalent to:

Ds3*)
shoot' $\wedge$ $\underline{\mathbf{M}(}$alive'') $\Rightarrow \neg$loaded'

There is not a formal recourse to obtain this new rule Ds3* based on Ds3. The more complex structure of Reiter's defaults is helpful in not producing causal asymmetries. Nevertheless from the epistemological point of view, based on Ds3 one could intuitively suppose that Ds3* is acceptable. This is the problem we need to face.

We can assume priorities but, anyway, we would need to distinguish causal and not causal sides of this asymmetry. With these poor formal conditions we do not have yet an adequate distinction between the elements of the causal asymmetry. We can try a distinction in the level of argumentation, as Geffner´s proposal. Consequently we could identify the adequate consequences of a set of defaults, given certain information. But, in this way, we would not make the distinction, except in a merely stipulative way, between both default reasonings contained in the asymmetry.

On the contrary, let's suppose a different internal structure of a default rule, such that it contains a formal structure instead the justifications. In a more general way, in [7], I called "proviso" to this new structure as a generalization of the particular structure of Reiter's justifications, and the name was assumed also in [8]. Such a new structure should verify certain propositional connections. If those connections were fulfilled the default would be applied. If they were not fulfilled, the default would be blocked. We will take the Shoot example to show this. This structure appear in [8] as an element to elucidate scientific explanation in a formal frame called ***GMD***. The structure implies to introduce certain other formal instruments like the following:

We say that:

(a) [chain$_n$] is a chain of linked phenomena, chain$_n$ is the set of elements of [chain$_n$], and D(chain$_n$), is the propositional description of the elements of chain$_n$.
(b) The set of edges of [chain$_n$] is $\{(e_1, e_2), (e_2, e_3),\ldots,(e_{n-1}, e_n)\}$.
(c) $e_w \nleftrightarrow$ [chain$_n$] means that e_w is the last linked phenomenon in [chain$_{n+1}$].
(d) D(e), D(chain$_n$), D([chain$_n$]), D(e $\nleftrightarrow$ [chain$_n$]) stand respectively for descriptions of some phenomenon e, a description of the elements of [chain$_n$], a description of [chain$_n$] itself, and a description of e_w being the last element in [chain$_{n+1}$].

With these notation we can represent a chain (as: [chain$_n$]) of phenomena e_i of a set of phenomena $\{e_1, e_2,\ldots,e_n\}$, the descriptions of the chain (as: D([chain$_n$])), and the connection, as a last element in the chain, of a particular phenomenon e_w in the chain (as: $e_w \nleftrightarrow$ [chain$_n$]). The central idea is to describe connections supposed in the world such that affect the default inference.

Even though these new elements in the structure of a default rule are used in ***GMD*** to a particular aim, such that other modifications to a default rule are implied, we can use the elements directly in a simple structure default. We will assume that [chain$_n$] is a **causal chain** of events. Therefore, we can add these elements in the following way:

(**cd**)

$$\text{shoot'} : \underline{\mathbf{M}}(\text{loaded'} \wedge e_w \nleftrightarrow [\text{chain}_n]) \Rightarrow \neg\text{alive''}$$

In this case the proviso is, then: $\underline{\mathbf{M}}(\text{loaded'} \wedge e_w \nleftrightarrow [\text{chain}_n])$.
We assume, first, that ¬alive'' = D(e_w), that is, e_w is the event referred to by ¬alive''. In the second place, we assume that D(chain$_n$) $\in E^h$ (an extension indexed with h, of the theoretical context). In the third place, we verify the proviso of **cd**. According to these assumptions, we can read **cd** as "If shoot' is derived in an extension E^h of a default theory $\Delta = <D, W>$, such that **cd** $\in$ D, and it is possible to assume the assertion loaded' and the assertion $e_w \nleftrightarrow$ [chain$_n$], in E^h; then, infer ¬alive". In natural language, the mean of **cd** is "If in my set of beliefs I have the description of a causal chain starting with the event of the shoot, and the occurrence of the shoot; then if it is possible to assume that the gun was loaded, and the description of the extended causal chain finishing with the event that the person in question is not alive; then, we can plausibly infer that the person in question is not alive.

Given the original example, and our new formal instruments, the last default rule should be equivalent to:

(**ncd**)

$$\text{shoot'} \wedge \underline{\mathbf{M}}(\text{alive''} \wedge e_u \nleftrightarrow [\text{chain}_p]) \Rightarrow \neg\text{loaded'}$$

We assume, first, that ¬loaded' = D(e_u), that is, e_u is the event referred to by ¬loaded'. In the second place, we assume that D(chain$_p$) $\in E^k$ (an extension indexed with k, of the theoretical context). In the third place, we verify the proviso of **ncd**. According to these assumptions, we can read **ncd** as "For some E^k of $\Delta = <D, W>$, if shoot' $\in E^k$, such that **ncd** $\in$ D, and it is possible alive'' and $e_u \nleftrightarrow$ [chain$_p$], in E^k; then, infer ¬loaded'.

With the last formal construction of default rules, we note that **cd** will be applied while **ncd** will not be applied. This is because the Default **ncd** is blocked for $e_u \nmid [chain_p]$ is not plausible. In standard contexts it is not plausible to complete the causal chain $[chain_p]$ linking finally with the event e_u. Thus, **ncd** fails to infer its conjecture because it fails in assuming certain propositional connections involved. This plausibility is relative to a particular context represented by Δ. In this way, we can establish a set of relevant connections among referents of propositions. So then, we can establish priorities among them. These relevant connections could sanction the priority among the elements of a set D of defaults. This procedure promises to make coherence between the argument perspective and its argumentation perspective, to face the problem of soundness in defaults. The soundness of default reasoning is a result of the fulfillment of corresponding provisos with its internal constraints, its propositional connections assigned. A good default reasoning is such that its ontological contextual constraints hold. The relationship expected is not only formal. On the contrary, in an extra-system sense, it is predominantly a semantic relationship.

Philosophically, the next problem to pursue is the problem of a relativistic approach of propositional connections. I think that the problem of context-dependence not necessary lead us to a relativistic position about the propositional connections. To solve this, in this proposal, I maintain the kind of propositional connection as **context-invariant**, even though their particular cases, i.e., even if the instances of the propositional connection, are **context-variant**. This is the centre of the formal proposal: system of beliefs can deal with fixed kinds of connections whereas the particular cases of those connections, i.e., the particular ordered pairs fulfilling the connection, vary from context to context. This strategy is clearly an intentionalist perspective of connections between events. Note that this strategy also avoids a relativistic approach[14]. This relativism is not desirable, given the epistemological position of analysis of the defaults developed here.

In this way we can assume the same propositional connection in the chain of both, **cd** and **ncd** default rules, under the same theory Δ. But, in a very different situation, we could assume two different default theories Δ1 and Δ2, both having the same notion of causality. This allows certain kind of comparability. The important difference between them is that Δ1 could assume different particular cases having the possibility of keeping that causal connection. If we believe in ghosts, and we believe that a ghost make a horrible sound, our idea of causal connection could be the same as that of the system of beliefs of a skeptical person about ghosts. She will believe, let's say, that there is a causal connection between the wind and a crystal bottle on the table. Nevertheless, our notion of causal connection does not need to be different than her notion of causal connection. The difference could be put in terms of what things can keep the causal connection in our ontologically different worlds. Maybe another person believes in ghosts. But this person may be more cautious than us and refuses to believe that ghosts causally generate something in the physical world. The propositional connection is invariant but the *relatas* supposed to fulfill it are variants.

[14] A first version of this intentional strategy, appeared in [7], but related with the discussion about scientific explanation.

Obviously, the last proposal is an intentionalist perspective about propositional connections. I am convinced that this position is the key in a characterization of the notion of default logical consequence in the old approach. In that approach, we need to specify the intuitions underlying that notion. With that aim, I will take into account our proposal of the connections structure inside a proviso of a default rule that link in a stronger way prerequisite and conjecture.

According to these propositional connections to structure the proviso[15], default consequence is based on possibility more than on necessity. But this possibility has to hold certain contextual constraints. The formality of default is intervening in the inference. We cannot infer the conjecture unless we verify, in a particular context, that the events referred could keep the propositional connections assumed in the proviso. In that sense, the inference depends on the fulfillment of the structure of the default. The truth of the prerequisite and of the proviso is being transmitted to the conjecture. However, contrary to the deductive case, this transmission of truth is not necessarily in virtue of the formality of the propositional elements in the default schema. The transmission of truth depends on ontological constraints of the context associated to the default schema (the default theory). The propositional connections in the proviso are formal and general structures, but the instances that could fulfill that formal structure are determined by the particular context in question. In that particular context certain instances are possible and other instances impossible, according to the additional inferences that would extend the original system of beliefs (the original default theory). We infer by default conjecturing from possibilities, but these possibilities are not unrestricted possibilities. One and the same default rule can be blocked in a theory Δ_n (in a given extension) and applied in a theory Δ_m (in a given extension). In the case of applied defaults, the prerequisite and the proviso are the information that triggers the inference that allows the conjecture. It could be the case that the information of the prerequisite and the proviso is not sufficient, in the sense of a deductive relationship of inference. But they transmit the truth to the conjecture in a fallible and retractable sense of transmission. It is not an infallible transmission of truth, it is not sufficient information to that aim, but it is also a transmission of truth. Maybe we could call this kind of transmission of truth "**transitory transmission of truth**".

5.2 Pretheoretical Definition of Default Logical Consequence

A default argument is sound respect to a particular context if in at least one way of inferentially extending that context it occurs that **If** (1) the prerequisite is true, and (2) the *relatas* to its propositional connections in its proviso are admitted as possible *relatas* of that propositional connections, **then** in that extension of that context the truth of the conjecture is transitorily assumable.

The success of this pre-theoretical definition depends on guaranteeing a stronger relationship between prerequisite and conjecture, by means of propositional connections demanded in the proviso.

[15] In [8] other possible modifications on the proviso are suggested. They are called "modulations".

5.3 Explanation of Default Logical Consequence

Within the framework of a default theory $\Delta = <\mathrm{D}, \mathrm{W}>$ in which the elements of D have propositional connections in the proviso linking prerequisite and conjecture by means of proviso, and in which we assume typical axioms of First Order Logic and corresponding rules preserving truth by their form; something is a sound default argument with respect to Δ if there is at least an extension E of Δ such that its prerequisite is in E, and its proviso is consistent respect to E.

Alternatively, we could characterize this definition in a slightly more formal way, using the notion of proof as following: Within the framework of a default theory $\Delta = <\mathrm{D}, \mathrm{W}>$ in which the elements of D have propositional connections in the proviso, linking prerequisite with conjecture by means of the proviso, and in which we assume typical axioms of First Order Logic (FOL), and corresponding rules preserving truth by their form; something is a sound default argument with respect to Δ if there is at least an extension E of Δ such that there exist a sequence $<\gamma_1, \ldots, \gamma_n>$ satisfying the following conditions:

(1) $\{\gamma_1, \ldots, \gamma_n\} \subseteq \mathrm{E}$,
(2) γ_n is the conjecture of the default, and is obtainable by the default at that stage of the sequence.
(3) For every γ_i in $<\gamma_1, \ldots, \gamma_n>$ occurs:
  (3a) γ_i is an axiom of FOL
  (3b) or γ_i is the conclusion of a rule of FOL, using $\gamma_\mathrm{k}, \ldots, \gamma_\mathrm{q}$, such that every number s in the interval [k-q] is such that s < i.
  (3c) or γ_i is the conclusion of a default rule **d** of D, using $\gamma_\mathrm{k}, \ldots, \gamma_\mathrm{q}$,. such that every number s in the interval [k-q] is such that s < i, and the proviso of **d** is fulfilled in E.

6 Conclusions

It have been developed an epistemic analysis of the interaction of the internal components of the default reasoning. I have started from the analysis of an argument as subject of study and not from an argumentation perspective. And I start the analysis taking default rules in the style of Reiter [1] as a model of our investigation. One of the most interesting result of that analysis was the essential disconnection between the internal components of a default, initially developed in [7]. Then I proposed an intentional perspective to understand the proviso of a default argument with the aim of guaranteeing the connection between prerequisite and conjecture. The kernel of that proposal is constituted by the notion of chain of referents [8] as a base. In this way we proposed the notion of context-invariant propositional connections associated to context-variant *relatas* for those connections. This allowed us to maintain contextualism in default reasoning without relativism. Then these formal resources were used in a definition of a notion of soundness in order to provide more precise intuitions in explaining syntactically the default logical consequence on the old approach.

I think that with these resources a solid advance can be made toward an integral solution of the problems of interaction and discrimination among elements of a set of defaults. I also hope that this epistemic and intensional proposal of reconstruction of default reasoning be useful to the investigation of the notion of logicality of non deductive and non-monotonic reasoning.

Acknowledgements. This research was benefited from the PAPIIT-UNAM project "Non Classical Logics and Argumentation in Science", IN401619; as well as from the discussions carried out in the Special Interest Group in Non-Monotonicity, *SIGNO-MON*, UNAM; and in the group of Formal Epistemology and Non Classical Logics of the Research Program in Science and Philosophy, *PRINCIPHIA,* UACM.

References

1. Reiter R (1980) A logic for default reasoning. Artif Intell 13:81–132
2. Reiter R, Criscuolo G (1981) On interacting defaults. In: Drinan A (ed) Proceedings of the seventh IJCAI conference, Vancouver, Canada, pp 270–276
3. Geffner H (1992) Default reasoning: causal and conditionals theories, 1st edn. MIT Press, Cambridge
4. Achinstein P (1983) The nature of explanation, 1st edn. Oxford University Press, Oxford
5. Van Fraassen BC (1980) The scientific image. Oxford University Press, Oxford
6. Gómez Torrente M (2004) La noción de consecuencia lógica. In: Orayen R, Moretti A (eds) Filosofía de la Lógica, vol 27. Enciclopedia Iberoamericana de Filosofía, Trotta, Madrid
7. Gaytán D (2014) Un modelo no monotónico y paraconsistente de explicación científica. Ph.D. thesis, Universidad Nacional Autónoma de México, UNAM, México
8. Gaytán D, D'Ottaviano IM, Morado R (2018) Provided you're not trivial. Adding defaults and paraconsistency to a formal model of explanation. In: Carnielli W, Malinowski J (eds) Contradictions, from consistency to inconsistency. Trends in logic, vol 47. Springer, Switzerland

Models and Data in Finance: *les Liaisons Dangereuses*

Emiliano Ippoliti(✉)

Department of Philosophy, Sapienza University of Rome, Rome, Italy
emiliano.ippoliti@uniromal.it

Abstract. In my paper, I examine the role that models play, and the relation between models and data, in financial systems, in particular in stock markets. I discuss several *dangerous liaisons* between models and data both from a theoretical and a practical viewpoint, and I consider their effects on the behaviour of the financial systems and their actors.

I will examine two issues in particular.

First, these *liaisons* defy the way traditional philosophy of science accounts for models and the relation between models and data, as stock markets exhibit several dynamics and features that do not fit them. For instance, they challenge the ontological issue about models (the debate about the fictional character of models), or the way models and phenomena are connected, and consequently undermine classical taxonomies as 'models of data', 'models of phenomena', 'models of theory'.

Second, these relations and liaisons open the way to possible exploitations and manipulation of stock market's dynamics by means of appropriate use of models and data, including '*reverse finance*', which should be closely analysed in order to contribute to better functioning and serving of the financial systems.

1 Introduction

As their use in financial markets has increased powerfully in the last few decades, models, and the role they play in finance, have been closely scrutinized and questioned.[1] One of the main concerns about them is the fact that "financial markets have become model-driven" (Svetlova 2012, 2), and this change has been so drastic that the markets "have nearly been destroyed in the course of this development. Models are considered an important contributing factor to the crisis; supposedly, they not only ignored the potential for the turmoil but also enhanced the crisis by miscalculating risks and mispricing assets" (Svetlova 2012). Thus, while financial models have been introduced and developed, especially by quants, in order to produce more 'objective' and 'accurate' descriptions, or predictions, of financial phenomena, it turned out that "the major problem seems to be that the financial market participants relied excessively on models and thus contributed to the crisis. When investors apply the same flawed formula, damage can arise" (Svetlova 2012).

[1] In particular see Boldyrev and Ushakov (2016), Brisset (2017a, 2016, 2017b, 2018, Brisset 2019), Svetlova (2012).

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 393–406, 2019.
https://doi.org/10.1007/978-3-030-32722-4_22

By 'model' here we simply mean a way of classifying certain objects, and their characteristics and relations (equivalence, implication, etc.), in a way that enables deductions. Thus, especially in finance, a model consists of an abstract world, or better 'credible worlds', and thus it is "a matter of categorization, assignment of relations and deduction" (Brisset 2019, 142). In more detail, a model characterizes a mathematical world the encapsulate a specific set of hypotheses. Under this respect, here we have a first problem, a first *dangerous liaison*, which is theoretical in kind: as Nancy Cartwright (2009) argues, a model embeds its environment, and the results that it gets are contained in its initial hypotheses. If this fact, maintains Cartwright, does not seem to be puzzling in physics, where the core hypotheses of models (the laws of physics) are considered to have a strong nomological power, things are very different in eco-nomic modelling. In the latter case, in effect, the application of the *ceteris paribus* clause becomes highly problematic and a model can easily fail to serve as a 'nomological machine'. To overcome this weakness, an extra level of justification is necessary, and it takes the form of a discourse backing the model—a discourse that makes the links between the hypotheses of the model and the specific phenomena under investigation as credible as possible. Thus, the external justification and the process of acceptance of a model can become very heaviy in economic modelling (and especially in finance). Now,

> the question of credibility, namely the legitimizing discourse of the formal model, refers directly to the conditions of felicity of scientific speech acts. Some procedures and contexts are socially accepted as indicators of the relevance of theories, including the refinement and robustness of modelling, falsificationism, the acceptance of the test procedure, the speaker's place in the scientific field and a number of elements directly involved in the immediate credibility of a theory. Thus, we see here that a theory is not reducible to its formal content, nor is it is a closed system. It is inevitably open to a social context bestowing its explanatory power upon it (Brisset 2019, 142)

Thus, the very inception and proper functioning of a model in finance is sensitive to the *discourse*, or the *narrative,* that back it. This discourse employs not only scientific lines of reasoning, but also rhetorical and persuasive devices in order to enable the model to gain consensus and credibility.

A second *dangerous liaison* is *practical* in kind, as the connection between the initial hypotheses of a model and the results that it gets in finance (that is, the connection between economics and economy) derives from the *application* of the model. This is a main tenet of the well-known 'performative' thesis, which received quite a bit of attention in the last few decades.[2] This thesis argues that when a model passes from the mere theoretical level to the *financial practice*, it affect the world it purports to describe: "the application of financial models not only informs and influences professional practices (generic and effective performativity) but also shapes markets by creating (or increasing) a fit between the real (market) world and the model world (Barnesian performativity)" (Svetlova 2012, 2).

The paramount example employed to back and substantiate the performative approach is the Black-Scholes option pricing model (BSM), which Donald MacKenzie

[2] See also Aspers (2007), Curran (2018), Davis (2006), Healy (2015), Vosslemand (2014) and Wullweber (2016).

notably uses to provide a historical example of a model that succeeded in shaping a market (i.e. the option market) after the majority of practitioners accepted it and started using it: "its use brought about a state of affairs of which it was a good empirical description […] practices informed by the model altered economic processes toward conformity with it" (MacKenzie et al. 2007, pp. 66–67).

MacKenzie's reconstruction of the development and employment of BSM, if right, shows that this model determines how option traders take their decisions: BSM is a device—a set of instructions—which market agents follow so that they obtain the same, or a similar, result (a price for an option). In this specific case, the performative power of the model is that if most of the traders calculate the option price for a given contract identically, then the market price would be the one given by Black-Scholes model. Bottom line: the model indeed produces the 'reality' that it describes, by aligning 'reality' to (the model) itself.[3] In other words, a model determines what happens: it 'makes itself true' and economy is taken over by economics—at least in part and temporarily.

If this is the case, that is, if the performative theory is confirmed, then we have at least three puzzles to consider carefully. First, from a more philosophical perspective, we have the problem of realism—the financial reality does not exist, it is simply constructed. Second, "if this theory was confirmed, then models would justifiably take centre stage in the debate on the causes of the crisis" (Svetlova 2012, 2). Third, the financial data we read on display are determined by the prevailing model employed by practitioners.

2 Making Markets and Performativity

The justification of the performativity theory is a well-known and highly debated issue in both philosophy and economics, and it is not a simple problem to solve since the concept of performativity is multifaced and has been used in a variety of ways.

The performative account of economics and finance originates from Michel Callon and Bruno Latour (1987)—and their Actor Network theory (ANT)—who, in turn, draw on Austin's account of performativity in linguistic (see Austin 1962, 1979). In a nutshell, the performative account states that "economics, in the broad sense of the term, performs, shapes and formats the economy, rather than observing how it functions" (Callon 1998, 2). The effectiveness of performativity, according to Callon, derives from its socio-technical nature, that is, the fact that markets are an assembly of humans and non-humans, or human beings and techniques/devices.

Now, in Austin's linguistics a *performative* utterance designates a speech act able to modify the world welcoming the enunciation, e.g. 'I pronounce you wife and husband' (A). Under this respect, this enunciation and a statement such as "GDP goes down" (B) seem to belong to two different registers of speech: (B) states something, (A) performs an action. Now, Austin's well-known thesis is that 'saying' is always

[3] More in detail, this means that, if the model underprices risks, we will have a reality (a market) where risks are underpriced; if the models misprice an asset, then that asset becomes mispriced in the 'real' markets.

'doing' something. This is the basis for the performative approach to economics: economists (an economic theory or model) do not just say, they produce the world they are meant to describe. Or better, describing the economic world *by itself* produces an effect on it.

On the same line of reasoning we find Donald MacKenzie:

> Economics is at work within economies in a way that is at odds with the widespread conception of science as an activity whose sole purpose is to observe and study, that is to "know" the world. The issue of that needs to be tackled in relation to economies and economics is not just about "knowing" the world, accurately or not. It is about producing. It is not (only) about economics being "right" or "wrong" but (also, and perhaps more important) about it being "able" or "unable" to transform the world (MacKenzie et al. 2007: 2)

Drawing upon Callon's ideas, Donald Mackenzie characterizes three kinds of economic performativity (MacKenzie 2006, pp. 16–18):

- *generic* performativity, which emerges when practitioners use a theory;
- *effective* performativity, which arises when using a theory produces effects on social reality.
- *Barnesian* performativity (Barnes 1983), which arises when the effects of using a theory align social reality to the assumptions or predictions of that theory—that is the theory becomes true, or self-fulfilling.

Of course, Callon and Mackenzie are not the first ones to theorize and acknowledge that economic theories do affect the course of economy and of the social world. For instance, Karl Marx (1844) maintains that the representations produced by economists are one of the engines of capitalism. In particular he argues that classical economic theory has actively shaped the phenomenology of the world so that it takes the form of the descriptions made by them.

Similarly, Polanyi (1977) argues that the theoretical model of the *rational man (homo economicus)*—a man motivated by self-interest and the desire for wealth—plays a crucial role in establishing the industrial capitalism, and thus in reifying the model itself. Real human beings progressively shaped themselves in accordance with the characteristics of the *homo economicus*, so making the model 'true'.

What is really at stake with the performative approach is the capacity to influence the social world in the direction of greater conformity to a specific theory or model. The concept of performativity is just a way to deal with this relation. It pushes further the influence of economics over economy and, accordingly, is highly controversial and has been strongly criticized. For lack of space, it is not possible to do justice to all the interpretations and applications of the concept of performativity, so here I will recall three main lines of criticism. The first one questions the concept and the meaning of performativity inside economics. The second one maintains that the theory of performativity is defective since it does not provide an account for performative inefficiency, that is, it does not identify "conditions of felicity"—conditions for success and failure of a theory or a model. The third one complains the lack of a real, cogent example of economic performativity, since BSM's case examined by Mackenzie would not be so.

The first line of criticism is put forward by Mäki (2013), who questions the performative thesis, in particular Donald MacKenzie's use of the concept of performativity. In a nutshell, Maki argues that:

(a). MacKenzie is employing a (wrong) concept, which is different than Austin's definition of performativity (and then we should "save Austin from MacKenzie", see Maki 2013).
(b). *Making a market* is not the same as to be performative (see Maki 2011).

Since performative speech acts are not simply true or false, but follow a self-referential logic that depends on conditions that are internalized by actors as truths, Maki notes that in order move from language to economics, we have to keep well separated illocutionary (that is, performative) and the perlocutionary acts—they are two different ways, constitutive and causal, for language to have an effect on reality. In more detail, Maki maintains that the Austinian definition of an illocutionary act produces a *constitutive* relation, while the perlocutionary a *causal* relation. A counter-objection to Maki is put forward by Guala (2016), who argues, building upon Searle (1995) and Millikan (2014) works, that the distinction of these two acts is not so easy to defend and that the institutional facts of which the acts of language are constitutive "are nothing but sets of actions and expectations about actions" (Guala 2016, 48). In this sense an act of language can be viewed as the "creation or maintenance (enforcement) of a set of beliefs connected to expected behaviors following an utterance: a promise will count as a promise only if it *causes* expected effects" (Brisset 2019, 140)

A way out to this puzzle is the one put forward just by Brisset, who argues that we should drop the distinction between causal and constitutive effects, in favor of one that employs the distinction between necessary and contingent effects: "the fact that the statement 'I do' implies that one regards me as administratively dependent on another individual is a necessary effect (illocutionary) of the speech act. Without these effects, we are not married" (Brisset 2019, 141). In more detail, such a solution aims at showing that the real debate is "between institutionally or conventionally controlled effects and uncontrolled effects. The fact of saying 'I do' causes a securing of my expectation over the behaviors of the other members of my community". In economics, and especially in finance, this means that "theories create new representations, new behaviors, which will lead to institutional changes. The consequence of such reasoning is that the frontier between illocutionary effects and perlocutionary effects is located historically, and thus is moving over time: perlocutionary effects can become illocutionary effects. This means that an unexpected effect can become an effect that must follow a felicitous speech act" (Brisset 2019).

Bottom line: performativity has to be intended, in line with Callon, as the "notion of 'make to act' through the intervention of a convention resulting from economic theory" (Brisset 2019, 142).

Once the concept of performativity has been clarified and better founded, it remains open question whether it does work in economics and finance. This leads us to the second line of criticism, that is, the issue of the 'conditions of felicity' of theoretical speech acts, and thus the question of performative failures. In effect, if on one hand it is hard to deny that performativity is not at work in economics, on the other it is hard as well to characterize how and to what extent it does work. It is plenty of cases of models or theories that do no 'perform' and "in the literature, we can find only one example of strong performativity, namely the Black-Scholes model" (Svetlova—Dirksen 2014, 562). The application of models in financial practice does not produce always performative effects, and

this "rarity of clear cases of Barnesian performativity is striking, so MacKenzie et al. (2007, p. 78ff) was forced to explain the uniqueness of the BSM in this respect" (MacKenzie et al. 2007). Of course, as noted by Butler (Butler 2010), the fact that a scientific model does not perform is not against the performative approach, but it requires to produce an account for "performative inefficiency" (Sotiropoulos and Rutterford 2014): there is something that resists or prevent performative processes, and a mature performative approach is expected to clarify the circumstances under which performativity can occur or not.

Of course, both Austin's and Callon's theories contain a (more or less explicit) answer to this question. After reconstructing these answers, Brisset (2019) shows us that performativity requires conditions of felicity to be increasingly present in order to take place and in his 'conventionalist' solution to the problem he identifies three conditions for performativity—beyond social and moral criteria or limits to the performativity of economic theories. These conditions enable us to answer three basic issues: (a) how a performed theory gains an empirical status necessary for its performance; (b) how does the theory in question auto-realize in the eyes of social actors; (c) how it is made compatible with conventions that structures the social environment.

In more detail, the three conditions or limits to the performative dynamics suggested are:

1. a limit of empiricity, that is, the fact that a theoretical product "can only be performed to the extent that it provides a point of reference for social agents (who may be politicians)" (Brisset 2019, 153) and "the choice of theoretical output which is to be elevated to the rank of empirical principle intended to be tested or applied to the economy can be explained in cultural terms corresponding to the vision of the world" (Brisset 2019)
2. a limit of self-fulfillment of the theory or model to perform: "in order to understand performativity, one is required to grasp either the way in which theory serves as an effective reference for agents or what it is that masks the dissonance between social beliefs and social phenomena" (Brisset 2019)
3. a limit of posed by the "cultural context of the output of economic knowledge" since the modal has to fit the constraint of conventional coherence" (Brisset 2019).

The third line of criticism to the performative approach argues that, in order to gain theoretical robustness, this has to be improved with the production of better case studies. In particular, this line argues that even MacKenzie's case study, that is, the BSM model, would not provide a real, uncontroversial example of performativity.

In effect Mackenzie contributed in a crucial way to look at the BSM as the stock example of a performative model, that is, a model that was employed by traders not to *discover* pre-existing price regularities, but to *anticipate* option prices in their arbitrages, so to generate option prices that correspond to the theoretical prices derived from the BSM model. Nevertheless, Brisset presents arguments that seem to support the idea that is not *completely* so, as "the BSM model never became a self-fulfilling model. I would claim that the October 1987 stock market crash is empirical proof that the financial world never fit with the economic theory underpinning the BSM" (Brisset 2018, 1). I will not enter this debate here, but it is enough to note that this does not mean, and Brisset does not mean, that the performativity is not at work here—and in

general in finance. The simple occurrence of 1987 October (the Black Monday) market crash does not imply that the BSM did not perform, *at least temporarily*, and it is hard to argue that the BSM did not play a crucial role in determining the prices of options before the black Monday.

In general, it is hard to deny or neglect performative processes in economics, and especially in finance. As well argued by Muniesa (2014), in these fields just testing a theoretical system does create the object that it theorizes. For example, the act of looking for a test of preferences for perfumes, requires creating a device that not simply record, but creates preferences: before applying this device, there is no order on the various perfumes.

Thus, in this sense, the performative process can be legitimately seen as a chapter of the classical epistemological debate on theory-ladenness of data and observations. Our analysis of a case study in algo-trading (see § 4) will provide further evidence to this contention.

3 Models and Data in Finance

Once a model or a theory has been produced, it goes without saying that in order to perform the market, it must be increasingly put in use by practitioners. The way and extent to which it is applied becomes essential in this sense: the "cultures of model use", in Svetlova's terms (Svetlova 2012), is central for its performative power. This point has been stressed also by Donald MacKenzie:

> The work on models by researchers on finance influenced by science studies (work that forms part of the specialism sometimes called 'social studies of finance') has focused primarily on the 'performativity' (Callon 1998, 2007) of models, in other words on the way in which models are not simply representations of markets, but interventions in them, part of the way in which markets are constituted. Models have effects, for example on patterns of prices. However, vital though that issue is – we return to it at the end of this paper –exclusive attention to the effects of models occludes the prior question of the processes shaping how models are constructed and used (MacKenzie and Spears 2014, 395).

In particular, a culture of model use is shaped by the way practitioners deal with the output of a model:

> However, attention to the issue is needed when analyzing modelling, because the cultural dope has a close relative: what one might call the 'model dope', the person who unthinkingly accepts the outputs of a model. Model dopes are invoked routinely in public discussion of financial markets, and in seminar presentations we have encountered audiences utterly convinced that they must exist, indeed that they are pervasive. However, empirically it is far from clear that model dopes do exist: the contributions to the nascent social-studies-of-finance literature on modelling that have addressed the issue have all failed to find them. Mars (1998; see Svetlova 2012) shows how securities analysts' judgements of the value of shares are not driven by spreadsheet models; rather, they adjust the inputs into these models to fit their 'feel' for the 'story' about the corporation in question. Svetlova (2009, 2012) finds similar flexibility in how models are used: they are 'creative resources', she reports, rather than rules unambiguously determining action. Beunza and Stark find the traders they study to be '[a]ware of the fallibility of their models and far from reckless', and fully reflexive: indeed, traders practise 'reflexive modelling', in which models are used to infer others' beliefs from patterns of prices. This should not surprise us, point out Beunza and Stark: 'why should we deny to financial actors the capacity for reflexivity that we prize and praise in our own profession?' (MacKenzie and Spears 2014, 395).

In addition, Svetolova's findings suggest that "the mathematical structure of the model, especially the number of free parameters, might be relevant for the model's ability to influence markets" (Svetlova 2012, 15). She notes that while a model that "requires the determination of (too) many free parameters", can see its performative power "undermined in the process of practical use", a 'simpler' model, like the Black-Scholes one, "with just one free parameter […] managed to develop its performative potential to its full extent, at least temporarily" (Svetlova 2012). Mathematical simplicity is a factor for performability.

The discourse backing a model, the flexibility in the use of its output, its mathematical simplicity (see also Chen 2017 on this point), all contribute to mold the culture of model use. In turn, this culture provides us with constructive mechanisms that enable performativity. In this sense, performativity can be seen as a chapter of the philosophical issue of the acceptance of a theory.

Therefore, to have a performative behavior, at least two steps have to be fulfilled. First, the production of a *narrative* or a *discourse* (scientific and even non-scientific in kind), which back a model and its acceptability and credibility. Second, the practical application of the model by a certain number of practitioners.

When these two conditions hold, the system can generate data that reflect the content of model, or better, a model behave in a way to produce data that are theory-laden, *at least temporarily*. Thus, we do not have data that 'take a picture' of the 'real' world, and models that are constructed from them but, on the contrary, the specific content of the model (its hypotheses) shapes the 'real' world in a way to produce certain kind of data (and not other). Therefore, in this case, data are not the starting point, that is, the base from which a model can be derived, but they are intertwined with the model, or are even simply the end point of the model that acts in background. This dynamic opens the way to another interesting liaison between 'world' and 'model' via data, and which we can label, using a Boldyrev and Ushakov's (2016) expression, 'adjusting the model to adjust the world'. This liaison is the result of the "constructive mechanisms of the models, i.e., […] the built-in elements that are designed to 'implement' the model by adjusting the reality so that the model's propositions become true" (Boldyrev and Ushakov Boldyrev and Ushakov 2016, 38). We can note that the built-in elements can also be obtained by means of laws, norms, rules (e.g. rules of order execution), as well as technological innovations.

If this is true, then this particular aspect of the relation between model and data adds a new dimension to discussion of models in finance, since "apart from being a representation, a model may also be considered as an instrument of social engineering" (Boldyrev and Ushakov 2016). In other word, not only also a non-representational model, that is, one "focused on possible effects and are not concerned with realism of properties or processes" (Boldyrev and Ushakov 2016, p. 43) or with (high) unrealistic assumption, can be informative, but we can make these assumption and the consequent effects 'true' by adjusting the 'world' via norms, rules, laws, so to align the world to the model. In this way, we can run on 'reality' several constructive mechanisms that enact the outcomes described by the model. Of course, we do not mean that *any* economic model can be built so to be implemented in reality, but we can substantiate a model by means of self-implementing technologies and constructive mechanism, once it has been accepted and employed.

Thus, the two initial liaisons, along with the other described now, can temporally shape a market, or a portion of a market, and even manipulate it within certain time-scales and scopes. The example in the next paragraph will illustrate well this dynamic.

4 Micro-performativity

An interesting example of the ways a model or a technical device can perform financial data, i.e. prices, can be found in O'Hara (2010). In more detail, she provided us with an interesting, actual, example of the way a quote of a share is shaped by the application of specific technical devices—algorithms. This algorithm provides us with a neat example of a self-implementing technology, and a model that performs itself over a mraket. This algorithm, as well as many others with which it interacts, encapsulates a model that aims at price discovery and at providing liquidity while looking for conditions to make profits.

Let us see how this algorithm operates.

Starting from a situation where 900 shares are traded at 7.50 and the bid drops to 7.20, an algo-seller, instead of entering an order to sell at the current offer of 7.50, submit and then immediately cancel an order to buy just above the bid. Now if there is an another algo waiting to trade, then it may respond by bettering its own quote. The seller then repeats the process, submitting and immediately cancelling the order. The algo responds as well and, accordingly, the quote goes up. Note that this process "can drive quotes up without any actual trading. At some point, this behavior can induce other algos to join, fearful that the market is moving away from them before they have executed their pre-determined amounts. The seller can then hit the limit at the now much higher bid price" (O'Hara 2010, 13). More in detail, O'Hara reconstructs the process that produces a higher quote:

> from a bid of $7.20, the dueling algos drive bid prices up to $7.36 in the space of one minute without there being any actual trades. At this point the real algorithm cancelled (due to an odd-lot execution elsewhere), and the artificial algorithm reset to the original bid. It then begins anew to submit and cancel quotes instantaneously, attempting to lure others. The "game" begins again at 9:42 when another trader (most likely an algo) enters the fray, the bid again begins to rise (all without any trade occurring), the algos move into trade before prices get away from them in this time interval, and so on throughout the day (O'Hara 2010)

Basically, the model that runs the algorithm generates a sequence of data (quotes) like this:

Algorithm A	Algorithm B
bid 7.21 and cancel	
	bid 7.22
bid 7.23 and cancel	
	bid 7.23
bid 7.24 and cancel	
	bid 7.25
bid 7.26 and cancel	
…	…

Visually, this dynamic produces a characteristic pattern of quotes as the one showed in Fig. 1.

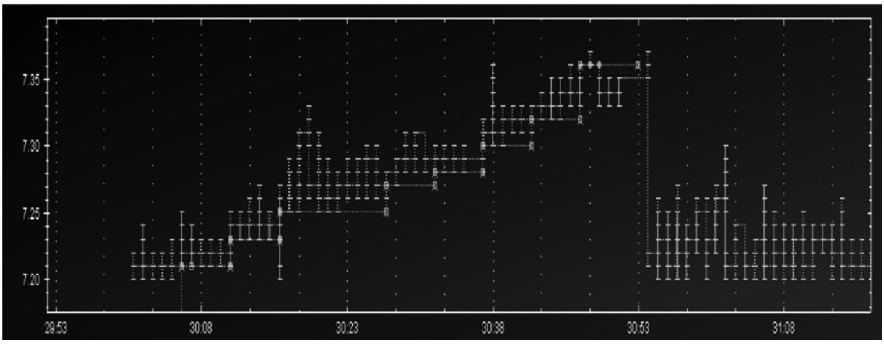


Fig. 1. The pattern of quotes generated by Algorithm A and Algorithm B

This sequence does describe a kind of performativity, for it temporarily shapes a market (or a portion of it) in accordance with a model that is programmed and run just to obtain a specific outcome (push the quote of a share up, and then sell it). Of course, the number of participants is essential to generate the (wanted) dynamics, just like one of the main tenets of performativity requires. Thus, the model approaches 'Barnesian' performativity as the number of players (other algorithms in this case) join it and follows the pattern originated by it. Of course, in order to perform, the model embedded in the algo has to meet certain circumstances (*conditions of felicity*), as O'Hara notes:

> while this strategy can be pursued in any stock with algorithmic trading, it is most likely to be successful for less-liquid stocks. This is because the high-frequency seller runs the risk that someone actually hits his limit order at the bid in the millisecond it exists before cancellation, forcing him to buy the stock. This outcome becomes more likely with greater order activity, suggesting that artificial quote-based manipulation strategies are a bigger problem for less actively traded stocks. This is also likely to be the case when artificial quotes are used to manipulate prices in crossing networks (O'Hara 2010, 14).

This model affects the market in a way that can be legitimately regarded as a case of *micro-performativity*: the simple act of putting in a bid (and cancelling it right after) triggers a sequence of reactions that change the price of a share *without a single a trade occurring*. Under given rules of orders execution, the micro-model temporarily shapes the market in the way anticipated by its set of instructions. As a consequence, the data we see are the direct outcome of the model, and the *liaison* model-data is straightforward. Of course, the model's capability to perform here is much more limited in time and scope: it produces effects (under specific conditions of felicity) at a much shorter timescale and more limited scope (i.e. micro) than broader models considered, for instance, by Svetlova or Mackenzie.

Nonetheless, this example raises a few philosophical issues.

First, we have an ontological issue: if this is a kind of micro-performativity, and given that it takes place without a single trade, then what is a quote by now? It is no more the simple expression of the matching of supply and demand, and then its ontological status is unclear.

Second, we have a issue in the *philosophy of financial practice*. If a quote does not reflect demand and supply, and it does not contribute to price discovery or liquidity, then, as suggest O'Hara, more non-displayed orders (or part of them) between the quotes should be exposed and have a minimum duration.

Third, we have also a manipulation issue, or better a connection between performativity and manipulation: "if the goal of posting quotes for milliseconds is only to take advantage of computerized trading programs or venues that rely on the NBBO, isn't that manipulation?" (O'Hara 2010, 16). In effect, if we can setup up micro-models able to perform themselves and temporarily align the market to them, we are creating the base for manipulation and possible exploitations of the dynamics generated in such a micro-performative fashion.

Moreover, this example provides us with other relevant aspects that are able to influence markets and, then, to perform (in addition to the ones listed in Svetlova 2012). A first feature is the timescale of the predictions of the models. The shorter the time, the greater performative power. Thus, the less the time-frame of the model, the more the data will reflect the content of the model. A second, more obvious one, is the role played by algorithms and the extent of their use, which can trigger courses of actions and then produce constructive mechanisms. The more a market is run by algorithmic model, the higher the chance to enact mechanism that perform action and affect data. In effect, these algorithms usually seek to exploit the rules of order execution of a particular stock market venue in order to pursue their aims and can temporarily perform data in a crucial way. These rules can affect the data in several ways (see also Ippoliti 2017 on this point): since the data become what the financial agents have to be act on, the understanding of rules and mechanisms of their composition and aggregation is essential for a understanding of financial systems and their evolution—a stock example in this sense comes from the study of *markets microstructures* (see in particular Preda 2017 on this point).

This leads us to another crucial liaison in the philosophy of finance and financial practice, that is, the one that involves *market design*. Since constructive mechanism and performative dynamics do provide us with theoretical tools to shape social interactions, market designers can devise rules, norms and laws in a way that push social actors to fit the image of the theory or a model, rather than *vice versa*. This, in turn, leads us to a kind of *reverse finance*, whereby we start from the model and then we try to produce the corresponding "reality", or better, we start from the wanted outcome, the one predicted by the model, and then we try to design or engineering as much as possible (with norms, rules, technologies) the hypotheses of the model needed to produce that outcome, so that the market will 'resemble' the model. Thus, the study of reverse finance is concerned with examining which hypotheses are necessary for getting specific economical results, which are the real starting point of the investigation, that is, we set a (wanted) goal and tries to find the hypotheses that, if satisfied, would produce it. In more detail, since the same result, say T, can be obtained from several sets of

hypotheses, a main aim of reverse finance is to find the minimal 'natural' set of hypotheses that is capable of getting T. In doing so, one can also show that specific hypotheses are necessary for T to hold, while other not such and can be replaced with other ones. By 'natural' here I mean the hypotheses, that is, initial conditions, that better fit a specific socio-economic environment at a given time.

5 Models and Data in Finance: Some Philosophical Remarks

The mentioned works and findings show how the role of models and the *liaison* between models and data in finance (in particular in stock markets) raise important philosophical issues.

First, their application defies the ways traditional philosophy of science accounts for models and the relation between models and data. In effect stock markets exhibit specific dynamics that do not fit them. On one hand, they challenge an ontological issue about models, namely the debate about the fictional character of models. The distinction between 'fiction' and 'reality', between models of a system and the system itself, short-circuits here: as we have seen, under certain circumstances we cannot even tell if a model is a fiction or not, or if 'reality' is simply the outcome of a 'fictional', 'non-representational', model. On the other hand, we can find ourselves in a position where we cannot tell if a theory or a model is good or not. It also turns out that in order to detect regularities and patterns, and foresee markets dynamics, we cannot count on simple quantitative approaches (for more on this point see also Ippoliti 2017a, b).

Second, the way models, data and phenomena are intertwined in financial systems undermines a classical taxonomy such as the one that tells apart 'models of data', 'models of phenomena', 'models of theory' (see Suppes 1966). As we have seen, the specific features of financial systems make it difficult to distinguish between data, phenomena and theory: sequence of data are theory(model)-laden, while a phenomenon can be the outcome of a non-representational, unrealistic, model.

Third, the dangerous liaisons between models and data, both theoretically and practically, open the way to possible exploitations and manipulation of stock market's dynamics by means of an appropriate use of specific models and data. In effect, since neither data nor models, at least under certain circumstances, are simply representations or description of a target system, the boundaries between performativity and manipulation can become very thin in certain circumstances, especially at micro-level and in stock markets, as we have seen. Thus, a convincing discourse, which can employ both scientific and non-scientific arguments, can be constructed to back the acceptance and then the employment of a given model (or a theory). This model is then employed and, as it gets increasingly used by practitioners, it performs specific dynamics which, in turn, shapes the data of the system, which will be altered in a way that is not visible, since they are simply outcome of it. These data, in turn, become the base for the action (or reaction) of part of the social actors, who might not be aware of the 'model-ladeness' of these data, and then they could be misled in several way, for instance by following a course of actions that can end up with a higher price for the ask side, or a lower price for the bid side.

A lesson we can learn from this is that not only data and models, but in general performativity, manipulation, market design or engineering, and what I have labelled *reverse finance,* are intertwined in subtle ways that deserve further investigation in order to contribute to better understanding, functioning and serving of financial systems.

References

Aspers A (2007) Theory, reality, and performativity in markets. Am J Econ Sociol 66(2):379–398

Austin JL (1962) How to do things with words. Harvard University Press, Cambridge

Austin JL (1979) Performative utterances. In: Urmson JO, Warnock GJ (eds) Philosophical papers. Oxford University Press, New York, pp 233–252

Barnes B (1983) Social life as bootstrapped induction. Sociology 17(4):524–545

Boldyrev I, Ushakov A (2016) Adjusting the model to adjust the world: constructive mechanisms in postwar general equilibrium theory. J Econ Methodol 23(1):38–56

Brisset N (2016) Performativity and self-fulfillment: the case of finance. Revue européenne des sciences sociales 54(1):37–73

Brisset N (2017a) On performativity: option theory and the resistance of financial phenomena. J Hist Econ Thought 39:549–569

Brisset N (2017b) The future of performativity. Œconomia, 7–3:439-452. https://journals.openedition.org/oeconomia/2746#quotation

Brisset N (2018) Models as speech acts: the telling case of financial models. J Econ Methodol 25(3):1–21

Brisset N (2019) Economics and performativity. Exploring limits, theories and cases. Routledge, London

Butler J (2010) Performing agency. J Cult Econ 3(2):147–161

Callon M (1998) Introduction: the embeddedness of economic markets in economics. In: Callon M (ed) The laws of the markets. Blackwell, Oxford

Callon M (2007) What does it mean to say that economics is performative? In: MacKenzie D, Muniesa F, Siu L (eds) Do economists make markets? On the performativity of economics, Princeton University Press, Princeton, pp 310–357

Cartwright N (2009) If no capacities then no credible worlds. But can models reveal capacities? Erkenntnis 70:45–58

Chen P (2017) Mathematical representation in economics and finance: philosophical preference, mathematical simplicity, and empirical relevance. In: Chen P, Ippoliti E (eds) Methods and finance. A unifying view on finance, mathematics and philosophy, pp 17–49. Springer, Heidelberg

Curran D (2018) From performativity to representation as intervention: rethinking the 2008 Financial crisis and the recent history of social science. J Theory Soc Behav 1–19

Davis A (2006) The limits of metrological performativity: valuing equities in the London stock exchange. Competition Change 10(1):3–21

Guala F (2016) Enacting the dismal science: new perspectives on the performativity of economics. Boldyrev I, Svetlova E (eds) Palgrave-MacMillan, New York, pp 29–52

Healy K (2015) The performativity of networks. Eur J Socio 56(2):175–205

Ippoliti E (2017) Dark data: some methodological issues in finance. In: Chen P, Ippoliti E (eds) Methods and finance. A unifying view on finance, mathematics and philosophy. Springer, Heidelberg, pp 179–194

Ippoliti E (2017a) Method and finance. A view from outside. In: Chen P, Ippoliti E (eds) Methods and finance. A unifying view on finance, mathematics and philosophy. Springer, Heidelberg, pp 3–15. ISBN: 978-3-319-49871-3

Ippoliti E (2017b) Method and finance. A view from inside. In: Chen P, Ippoliti E (eds) Methods and finance. A unifying view on finance, mathematics and philosophy. Springer, Heidelberg, pp 121–128. ISBN: 978-3-319-49871-3

Latour B (1987) Science in action: how to follow scientists and engineers through society. Harvard University Press, Cambridge

Maki U (2013) Performativity: saving Austin from MacKenzie. In: Karakostas V, Dieks D (eds) EPSA11 perspectives and foundational problems in philosophy of science. Springer, Cham, pp 443–453

Mäki U (2013) Performativity: saving austin from MacKenzie. In: Karakostas V, Dieks D (eds) EPSA11 perspectives and foundational problems in philosophy of science. Springer, Dordrecht, pp 443–453

Mäki U (2011) Economics making markets is not performativity. Draft

MacKenzie D (2006) An engine, not a camera. How financial models shape markets. MIT Press, Cambridge (2006)

MacKenzie D, Muniesa F, Siu L (eds) (2007) Do economists make markets? On the performativity of economics. Princeton University Press, Princeton

MacKenzie D, Spears T (2014) 'The formula that killed Wall Street': The Gaussian copula and modelling practices in investment banking. Soc Stud Sci 44(3):393–417

Marx K (1844 [2007]) Economic and philosophic manuscripts of 1844. Dover Publications, New York

Millikan R (2014) Deflating socially constructed objects: what thoughts do to the world. In: Gallotti M., Michael J (eds) Perspectives on social ontology and social cognition. Springer, New York, pp 27–39

Muniesa F (2014) The provoked economy. Routledge, London

O'Hara M (2010) What is a quote. J Trading 5(2):10–16

Preda A (2017) The sciences of finance, their boundaries, their values. In: Ippoliti E, Chen P (eds) Methods and finance. A unifying view on finance, mathematics and philosophy. Springer, Cham, pp 151–167

Searle J (1995) The construction of social reality. The Free Press, New York

Polanyi K (1977) The livelihood of man. Academic Press, New York

Sotiropoulos D, Rutterford J (2014) Performativity and financial markets: option pricing in the late 19th century. Innovation, Knowledge and Development Research Centre, The Open University

Suppes P (1966) Models of data. Stud. Logic Found Math 44:252–261

Svetlova E (2009) Theoretical models as creative resources in financial markets. In: Jansen SA, Schröter E, Stehr N (eds) Rationalität der Kreativität. Spinger, pp 121–135

Svetlova E (2012) On the performative power of financial models. Econ Soc. https://doi.org/10.1080/03085147.2011.616145

Svetlova E, Dirksen V (2014) Models at work—models in decision making. Sci Context 27(4):561–567

Vosslemand E (2014) The 'performativity thesis' and its critics: towards a political ontology of management accounting. J Acc Bus Res 44(2):181–203

Wullweber J (2016) Performative global finance: bridging micro and macro approaches with a stratified perspective. New Polit Econ 21(3):305–321



How Can You Be Sure? Epistemic Feelings as a Monitoring System for Cognitive Contents

Sara Dellantonio[1(✉)] and Luigi Pastore[2(✉)]

[1] University of Trento, Trento, Italy
sara.dellantonio@unitn.it
[2] University of Bari, Bari, Italy
luigi.pastore@uniba.it

Abstract. We explore the view that subjective experiences, termed *epistemic feelings* in the literature, accompany reasoning, intuition and other cognitive processes. These epistemic feelings are considered to have a wide range of functions, providing us with information about many aspects of our cognitive content. For instance, they can tell us whether something we know is certain, uncertain, interesting, boring, doubtful, ambiguous, correct, plausible, informative, relevant, coherent or related to other cognitive content. To understand how epistemic feelings work is therefore essential in comprehending any kind of thinking process, including any kind of reasoning process. In this work, we first analyze epistemic feelings with the aim of describing their nature and functions. Secondly, we explore the analogies and differences between somatic, emotional and epistemic feelings and try to show that epistemic feelings are not the same as epistemic emotions and that although somatic, emotional and epistemic feelings all make use of the same kinds of signals, they are triggered by different monitoring systems. Finally, we show that epistemic feelings guide us to choose specific kinds of reasoning processes and determine whether we rely on intuitions rather than on analytic reasoning. Finally, although epistemic feelings do not necessarily lead us to make the best choice in every context, they can be trained to provide us with better guidance.

1 Introduction

Someone is trying to persuade you that green herbs can cure several diseases because green is the color of positivity: you immediately *sense* a lack of relevance in the argument that makes it seem *odd* to you. On the other hand, when your friends tell you that their child skipped school every day this year and that he will therefore fail, you consider this conclusion to be *natural* and *obvious* and feel *very certain* that this is going to happen. If you are reasoning about the political situation in Europe, you might conclude that, over time, right wing parties will *probably* reach even more consensus; nevertheless, you might not feel *sure* about this.

S. Dellantonio, L. Pastore—This contribution is fully collaborative. The authors' order is alphabetical.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 407–426, 2019.
https://doi.org/10.1007/978-3-030-32722-4_23

What is it that makes us feel a conclusion is the natural and obvious consequence of a premise, or more generally that makes us feel any cognitive content is certain vs. uncertain, plausible vs. implausible, and so on? In the view we embrace, reasoning, intuition as well as other cognitive processes are accompanied by subjective experiences – i.e. by feelings – that are related to them. In the literature these are called *epistemic feelings*, or sometimes referred to as *noetic* or *metacognitive feelings.*[1]

Feelings can be roughly defined as subjective conscious experiences of various kinds characterized by a specific phenomenal quality: some of our feelings signal (make us aware of) *specific bodily states* like pain, hunger or cold, others signal (make us aware of) *specific emotions* or *moods*, yet others signal (make us aware of) specific *epistemic stances* such as whether we are certain or uncertain of a conclusion or whether an argument feels reasonable or not. Many other kinds of feelings are reported by the literature, but they are not directly relevant for our analysis.

Epistemic feelings are considered to have a wide range of functions and provide information related to the epistemic assessment of any cognitive content we possess. For instance, they tell us whether something we know is certain, uncertain, interesting, boring, doubtful, ambiguous, correct, plausible, informative, relevant, coherent or related to other cognitive content. Moreover, they also include the so-called feeling-of-knowing which tells us whether we know something even if we cannot retrieve it right now (cf. e.g. Koriat 2000), as well as many other specific feelings such as the tip-of-the-tongue phenomenon, the feeling of forgetting, the sense of how easy or difficult it will be to accomplish a certain task or to retrieve certain information, the sense of familiarity we have about certain information we cannot specifically retrieve, the sense of correctness and similar phenomena (for a brief review cf. e.g. Michaelian and Arango-Muñoz 2014; Proust 2015).

To understand how epistemic feelings work is an essential element in comprehending any kind of thinking process including reasoning processes and rationality. In this work, we first analyze epistemic feelings with the aim of describing their nature and functions. Secondly, we explore the analogies and differences between somatic, emotional and epistemic feelings and try to show that although somatic, emotional and epistemic feelings all make use of the same kinds of signals, they are triggered by different monitoring systems. Finally, we show that *epistemic feelings guide us to choose specific kinds of reasoning processes over others*: that they determine whether we rely on intuitions rather than on analytic reasoning; and, while they do not necessarily lead us to make the best choice in any given context, they can be trained to provide us with better guidance.

[1] Epistemic feelings concern the subjective experience we have of our thinking processes. They tell us how we feel about them. Being subjective, they do not tell us anything about whether our cognitive processes are justified from an objective point of view. For this reason, since epistemology typically concerns the justification of cognitive processes, the word 'noetic feelings' would be more appropriate. In fact, the word 'noetic' indicates the subjective element of experience, thereby underscoring the point that these feelings result from a subjective assessment of our knowledge. As will become clear in the article, the word 'metacognitive' indicates the fact that these feelings are about our cognitive processes and tell us something about them. In this article, we use 'epistemic feelings' because this is the term used in most of the interdisciplinary literature on this topic.

2 The Epistemic Assessment of Cognitive Contents

"Green herbs can cure several diseases", "This year my friends' son will fail", "Over time right wing parties will reach even more consensus", "That's where my phone is! I left it in the dentist's office!", "To get a university degree does not guarantee you will have a better job", "The butler is the murderer". These are all examples of epistemic contents we might be confronted with in our everyday life. These contents might be the conclusion of some sort of thinking process we carried out on our own or they might be the conclusion reached by somebody else with whom we do or do not agree. We might have come to the conclusion that *green herbs can cure several diseases*, but not be very sure about this. Or somebody else might insist that *green herbs can cure several diseases* but we might think that this is false and feel very sure about this. We might establish that *having a university degree does not guarantee you will get a better job* and yet not be completely certain about this. Or somebody else might tell us that *getting a university degree does not guarantee you will have a better job* and we might think s/he is wrong, but feel uncertain about this. After seeing half of a movie, we might conclude that *the butler is the murderer* and feel very certain about it. Or somebody else might suggest that *the butler is the murderer* and we might think s/he is right and be completely certain about it.

When cognitive or propositional content like this arises in our minds, first we assess its truth-value. That *green herbs can cure several diseases* or that *having a university degree does not guarantee you will get a better job* are statements that we may hold to be true or false. However, once we reach a conclusion, its truth or falsity is not the only epistemic information we have access to. Cognitive contents not only appear to us as true or false, but are also always accompanied by further 'clues' regarding our epistemic attitude towards this content. When we conclude e.g. *that the butler is the murderer*, we are not only aware that we hold this to be true, but we are also immediately aware that we are certain about it. In addition to that, we might also experience this conclusion as obvious, predictable and uninteresting.

Where do these clues regarding our epistemic attitudes towards cognitive contents come from? This question can assume two possible forms. On the one hand, we might be interested in understanding why in certain cases something appears to us as certain vs. uncertain, as expected vs. surprising, and so on: for example, we might want to establish why we are certain that the butler is the murderer, while we are uncertain that our football team will win the game next Sunday. On the other hand, we might instead wonder how we obtain the information that a content present to our mind has specific epistemic features such as being certain or surprising. Using the example mentioned before: We know we are certain that the butler is the murderer, but *how do we know that we are certain*? By what channel are we informed of this certainty? In this work, we will try to address both questions, even though we are mostly interested in the latter.

To begin, let us start with an example. People working in scientific fields usually think that all healthy children must be vaccinated, and that autism is not caused by vaccination. When you have this belief, typically you do not only hold it to be true in a 'cold' manner. Most often, you also have a strong feeling of certainty. (Unfortunately, the same strong feeling of certainty also accompanies the opposite conclusion for

people who belong to the anti-vaccination party). Vaccination is a very clear example of the fact that – when faced with a given cognitive content – we are not only aware of the truth-value assigned to it by our thinking process. At a minimum, when we hold a conclusion true or false, we also know whether we are certain or uncertain about it.

Moreover, being certain or uncertain about our conclusions does not take the form of propositional information such as "I am very sure!". We do not become aware of being very sure (e.g. that vaccination does not causes autism) through a neutral, cognitive assessment, and we do not need any explicit or deliberate reflection to know we are sure. What informs us about our certainly is not so much explicit cognitive content additional to the conclusion we have already reached, but rather is a 'hot', non-propositional and immediate feeling, i.e. a certain sensation of confidence and drive. This also has a motivational effect, e.g., we may fight for our belief (on this cf. e.g. Arango-Muñoz 2014).

Being certain or uncertain, or being sure we are right or wrong are not the only feelings we experience in relation to our cognitive processes. There are a number of other feelings that can characterize our epistemic life and accompany our cognitive processes. As already mentioned, when we conclude, for example, that the butler is the murderer, in addition to the feeling of being certain about it, we also experience the feeling that it was *obvious*, *predictable* and that the story is *not interesting*. Thus, we have a 'hot', non-propositional and immediate feeling that informs us of the fact that the plot of the story is expectable and trivial.

Over the last years, an increasing number of studies have suggested that feelings inform us about a number of the epistemic attitudes we experience. These are called *epistemic feelings*. A list of examples, often mentioned in the literature, relevant to understanding what epistemic feelings are and how they work include (for a brief review of these feelings cf. e.g. Michaelian and Arango-Muñoz 2014; Proust 2015; de Sousa 2009):

- the *feeling of familiarity* or the *déjà-vu experience* that occurs when you have the impression you have already seen someone/something, even though you do not recognize anything specific about them;
- the *feeling of knowing* and the *feeling of having a name on the tip of your tongue*, i.e. the feeling that we do know something/a name, even though we cannot retrieve it right now;
- the *feeling of understanding* that we experience when we (finally) get the point of something;
- the *blank mind experience* when we feel that there is no content that can be brought to our consciousness right now;
- the *feeling of easiness or difficulty* we experience when we are carrying out a cognitive task, which is also predictive of the probability we will complete it and do so adequately;
- the feeling of *wonder*, *curiosity,* or *doubt* which motivates us to inquire more deeply;
- the *feelings of relatedness/unrelatedness* we have when different pieces of information are put together in the wrong manner (e.g. green herbs can cure several diseases because green is the color of positivity);

- the *feeling of error* or of *correctness* we experience while doing a cognitive task, when we sense that something we did was wrong or incoherent and we should check it again or, conversely, that something we did is correct and coherent and we can be happy about it;
- the *feeling of forgetting* that we experience toward certain situations when we sense that something we did – e.g. a certain course of action – was incomplete, but we do not know exactly which elements were missing;
- the *feeling of clarity* or *ambiguity* that characterizes, for example, a text, an explanation or a story we hear;
- the *feeling* that something is *interesting* or *informative* or *relevant* (or the opposite).

These are only some cases of the feelings we experience in our lives. Many of them certainly feel familiar (!) to most of us. They accompany our cognitive processes and give us further information on their epistemic status. These feelings are produced automatically, without any reflection. Even though they are not propositional, they convey information on the cognitive processing they accompany (Schwarz 2010). In fact, they give us feedback on our cognitive processes and for this reason they are often also called metacognitive feelings. This feedback is perceptual rather than propositional and allows us to assess our cognitive processes from various epistemic points of view, e.g. certainty, rightness, clarity, interest, etc.

Epistemic feelings are about various kinds of cognitive processes. The examples of certainty/uncertainty, being right/wrong given at the beginning of this section concern specific *propositional content*s (e.g. 'Green herbs can cure several diseases', 'This year my friends' son will fail', 'Over time right wing parties will reach even more consensus', 'Autism is not caused by vaccination'). These might result from an intuition we had or from a reasoning process we carried out or they could be conclusions reached by other people and presented to us. However epistemic feelings are not only about propositional contents like these. Feelings of familiarity accompany, for example, the *perceptions* we have of specific people, objects or places and tell us whether we have already seen them before (and how often). A feeling of easiness or difficulty can accompany *any kind of task* we are presented (whether it requires only cognitive resources or also involves some physical activity): both a mathematical calculation and a hurdle race can be perceived as easy or difficult. These feelings occur before we actually carry out the task as the effect of an implicit and automatic anticipation of their easiness or difficulty. They allow as to predict whether we will be able to do the task, how rapidly we can do it, and how much effort this will require. Another feeling that applies to any kind of task, whether it is purely cognitive or also practical, is the *feeling of having done things wrong* or *correctly*. After carrying out a task, we are not aware of all the steps we took (all the actions or cognitive operations we performed). And yet at the end, we might have the sensation that we did something wrong. This sensation is also what triggers an explicit recall of all the steps of the process as we try to identify the error. While you are typing a text, you might, for example, have the feeling that you made a mistake: this feeling motivates you to stop writing and go over what you wrote in order to look for the mistake.

The feeling of knowing concerns *memory recall performance*: it tells us whether specific information is available to us *without* – or, more specifically, before – *getting*

access to it. Michaelian and Arango-Muñoz (2014, p. 99) offer the example of participants in a TV quiz show: sometimes they press the buzzer not because they have already retrieved the answer, but just "relying on a gut feeling that tells you that you'll easily be able to retrieve it." And yet it is possible that this feeling is wrong and when they actually try to gain access to the content they are looking for, they will be unsuccessful. In this case, they experience the so-called tip-of-the-tongue phenomenon. The difference between the feeling of knowing and tip-of-the-tongue feeling is that the first occurs before we try to recover the information explicitly, when we do not yet know whether we will actually be able to retrieve the content we are looking for or not. By contrast, the latter occurs when we actually try to recover information we feel we have but are not yet able to retrieve. A feeling that is in certain respects analogous is the feeling of forgetting: imagine you are at the airport and suddenly have the feeling you forgot something. At the moment you get this feeling, you are not thinking about what you did before leaving the house, but because of this feeling, you start explicitly and deliberately trying to retrace your steps in order to recall what you forgot or whether you actually forgot something.

These last examples reveal an interesting aspect of epistemic feelings: *they can be about a number of different cognitive activities; their common denominator is that they can tell us something about underlying processes or contents we do not have conscious access to*. We do not have access to all our knowledge at any time (access consciousness is limited), but our feeling of knowing can tell us whether a content is or is not there before we actually search for it, while the tip-of-the-tongue experience tells us we are not able to retrieve a content right now, but is there and will pop out sooner or later. We do not actually remember everything we have done in a given situation, but our feelings of forgetting can tell us that an element in that sequence of actions may be missing. An analysis of these feelings – of their nature and functioning – could be of great help in understanding the way we think and reason.

3 The Nature of Epistemic Feelings: A Comparison with Emotional Feelings

In general, feelings *are brought about by bodily changes* and are *conscious experiences characterized by a phenomenal quality* that tells us something specific about the condition that gave rise to them. The simplest kind of feelings we all experience are those that depend directly on bodily changes and signal specific states of the body like hunger, thirst, cold, pain, muscle tension, etc. The function of *somatic feelings* is to inform us about the nature of the conditions that caused these states, e.g. stomach is empty, blood sugar is low, viscera are contracting in an excessive manner, tissue is lacerated, etc.

These somatic feelings are produced by our proprioceptive and interoceptive systems, which map the internal states of the body and keep track of all relevant changes. More specifically, the proprioceptive system oversees the position of the body and its parts (posture, muscle tension and position of the joints: Berthoz 2000, chap. 2), while the interoceptive system detects the physiological condition of the body (pain, temperature, itch, visceral sensations as well as sensations coming from the internal organs,

vasomotor activity, hunger, thirst, and respiration: Craig 2003, 2009, 2010; Ceunen et al. 2016). Thirst and pain are, for example, mapped by the same system, but they do not give rise to the same feeling: they are qualitatively different, and this is the reason why we do not confuse the different states they point to. Thus, feelings are not indeterminate; they are instead specific, and it is this qualitative specificity that makes them informative. When, for example, one feels thirsty (has a feeling of thirst) or feels pain (has a feeling of pain), s/he has the conscious experience of a sensation characterized by a specific quality. It is the specific quality of this experience – the way it feels – that informs the subject about the state of his/her body, in this case, about the need for liquids (thirst) or the laceration of the tissue (pain).

The quality of feelings can be determined, at least roughly, along various dimensions. First of all, feelings are characterized by a specific intensity. In the case of somatic feelings, the intensity (e.g. of the thirst or pain) informs us about the urgency and seriousness of the situation. Secondly, feelings have a valence (i.e. a hedonic tone) and can be more or less pleasant or unpleasant (pain is certainly unpleasant, while the sensation of feeling full is pleasant, at least if it is not too intense). Thirdly, feelings can be more or less short-lived or long lasting. Fourthly, they are affective in the sense that they 'affect' (influence) our cognitive processes and therefore our subsequent behavior, motivating one course of action instead of another (cf. e.g. Michaelian and Arango-Muñoz 2014; Proust 2015). Finally and more importantly, they are reactive and result from an appraisal of the situation. The notion of 'appraisal' describes the evaluation of whether and how a situation affects us and our well-being. Somatic feelings tell us, in particular, whether the situation that brought them about is positive or negative for us and in which respects. In the case of somatic feelings, appraisal is quite simple and is conveyed mainly by localization, and by the intensity and the hedonic tone of the feeling: a strong pain in the leg (localized in the leg, intense and with a negative hedonic tone) informs the subject that the situation is negative and caused by some problem in a specific part of the body. More generally, somatic feelings may be the product of a monitoring system whose task it is to keep track of all changes in states of the body and to evaluate them from the point of view of their import, urgency, valence and localization. The object of a somatic feeling is the bodily state it signals.

In some cases, bodily feelings are likely to be blends of feelings. Think, for example, of thirst. Thirst is characterized by the presence or absence of specific concomitant sensations. For example, the thirst we experience after a long run is characterized by the presence of specific concomitant sensations related to the muscles or the skin. These are absent in other kinds of thirst: they do not accompany, for example, the thirst we experience after having a hearty and too salty meal, which might be accompanied by different concomitant sensations in the stomach and in the mouth. In this example, the presence/absence of concomitant sensations helps identify the cause of thirst at a certain moment and is the basis for distinguishing between different tokens of thirst. However, it is plausible that in some cases these features may also be used to classify different types of states. In the case of 'hunger' vs. 'starvation', for example, some concomitant global sensation of extreme weakness might be crucial in determining that the state we are experiencing is 'starvation' and not just hunger.

Somatic feelings are very simple since they arise only in relation to detected bodily changes (above a certain threshold). In spite of their simplicity, in some cases they

might still be misleading. Feelings can never be actually wrong: they are just a signal triggered when a process results in a certain outcome. For example, when the interoceptive system registers specific bodily changes such us empty stomach or low blood sugar, the feeling of hunger is activated. In this case, the feeling is appropriate since it properly indicates the need of food. However, in some circumstances the feeling of hunger might also be triggered by bodily changes that are not directly related to the need for food. People suffering from diabetes often experience, for example, a symptom called 'polyphagia', i.e. an excessive hunger brought about by their condition. In this case, people experience the feeling of hunger even when they have already consumed more calories than they need. Thus, the feeling is correct (they actually feel hunger), but the process that triggers this feeling is anomalous since it is due to the pathology rather than a real need for food. In other cases, it is possible that feelings are misinterpreted: I might feel anxious but interpret this feeling as hunger. In this sense, although the feelings are always correct (if and only if a process results in a certain outcome, we experience a feeling), feelings can be misleading. In spite of all these caveats, feelings are clearly essential for us: to be aware of the states of our body is indispensable for our survival.

Yet, somatic feelings are not the only kind of well-known feelings. Other feelings with which we are well acquainted in our everyday experience, and which have been extensively studied by psychologists and philosophers, are the so-called *emotional feelings*, i.e. the kinds of feelings we experience during an emotional episode.[2]

According to Antonio Damasio, emotions are essentially brain processes characterized primarily by "automated programs of actions concocted by evolution" and secondarily by "a cognitive program that includes certain ideas and modes of cognition" (2010, p. 116). This view fits quite well with Paul Ekman's conclusion that at least some primary emotions called "basic emotions" are "affect programs", that is, specific evolutionarily determined brain mechanisms that control behavior and determine what stimuli cause what responses (see, e.g. Ekman 2003; on basic emotions see also Ekman 1992a, b; 1999). Basically, affect programs are automatic appraisal mechanisms: they work very fast, unconsciously and unreflectively and determine how we will react to minimally interpreted stimuli. To put it roughly, when we come into contact with the appropriate stimuli, emotion triggering brain regions are activated (such as the amygdala and frontal lobe areas) and signals are sent to the body; these signals cause specific bodily changes and trigger specific actions (including postural changes and facial expressions) as well as specific thoughts. These stimuli are also the object of the emotion, i.e. what caused it. Take the example of fear: our brain reacts to a specific stimulus, i.e., to the object of the emotions and what caused them (for example, a snake); this object is appraised as dangerous and brings about a number of bodily

[2] Our description of the feeling of emotion relies on a specific perceptual theory of emotion which we have extensively defended elsewhere and that reaches a compromise with cognitive theories of emotions: cf. Dellantonio and Pastore (2017), chap. 5. People who embrace the position that emotions are forms of cognition, i.e. beliefs, would also interpret epistemic feelings as second-order beliefs (e.g., I believe that autism is not caused by vaccination and I believe I am sure that autism is not caused by vaccination). For a criticism of this doxastic account of epistemic feelings, cf. Arango-Muñoz 2014.

changes (increases in heart rate, blood pressure and respiration, contraction of muscles in the gut, the face, etc.) and a specific (re)action ensues, which in this case might be a fight-or-flight response.

Secondary (cultural and social) emotions are more complex, require high level cognitive resources and are individually and culturally variable. However, even though they are neither universal nor innate or biologically preprogrammed, they are also appraisal mechanisms. The difference between basic and non-basic emotions is that the appraisal mechanism does not work automatically for non-basic emotions, requiring cognitive mediation for the evaluation of the stimuli and an interpretation of how the situation might affects the subject (they tell us that some element or situation in the external environment is relevant for our goals and whether it may have positive or negative implications for these goals – cf. e.g. Lazarus 1991; Frijda 1986; Oatley and Johnson-Laird 1987; for an overview of appraisal theories cf. Scherer et al. 2001). We might be angry that someone tried to break into our apartment: to get angry we need, first of all, to interpret the stimulus (the attempt to break in) as negative for us. This appraisal of the situation generates emotions which result – like basic emotions – in a number of bodily changes and dispositions to act.

Emotions that arise in social contexts are often complex not only in the sense that the situation that caused them needs a cognitive interpretation, but also because they do not occur in a pure form. Consider the case of the attempt to break into our apartment: even if the situation is quite simple, we will probably not experience only anger, but rather a blend of emotions that involves further cognitive elaborations and also includes disapproval, demoralization, worry, fear, preoccupation. Some emotions are in themselves (cognitively elaborated) blends of other emotions. Disapproval, for example, includes, among others, emotions such as sadness and surprise (for more on emotional blends and the role of cognition, cf. e.g. Lane and Schwartz 1987; Prinz 2004, chap. 4). Like basic emotions, cognitively elaborated emotional blends also give rise to a (complex) appraisal of the situation, which, in turn, brings about a large amount of bodily changes and dispositions to act.

The bodily changes that result from appraisal play an essential function: they give rise to specific feelings that inform us that we are experiencing an emotional episode. In Damasio's words: "emotional feelings are mostly perceptions of what our bodies do during the emoting, along with perceptions of our state of mind during the same period of time." (Damasio 2010, p. 117) Conversely, the channel through which we perceive those brain processes known as "emotions" – i.e. the way in which we come to know our emotions – are the feelings brought about by the complex of bodily (e.g. endocrine, cardiac, circulatory, respiratory, intestinal, dermal and muscular) changes that occur as a consequence of such brain states.

Since they originate from bodily changes and thus ultimately from the interoceptive and proprioceptive systems, emotional feelings are analogous – but, as we will see, not identical – to somatic feelings. It is therefore not surprising that they share comparable features.

First of all, emotional feelings are also characterized by specific qualities and in some cases such as, for example, fear and happiness, we can say they feel different to us. Yet, there is a long-standing debate about whether the qualities of emotional feelings are sufficiently differentiated from a perceptual point of view to be used to

discriminate among emotions, i.e., about whether we can identify the emotion we are experiencing on the basis of these qualities. Perceptual and cognitive theories have different views on this; however, a number of authors agree on a compromise position (for an overview of this debate and of the arguments in favor of this compromise position cf. Dellantonio and Pastore 2017, p. 223 ff.): emotional feelings constitute the perceptual information we use to identify at least basic emotions. In the case of complex emotions (cognitively elaborated blends of emotions), these feelings might be composite and amalgamated; thus, they might not be sufficiently determinate to be used as the *sole* means for identifying the specific emotions we are experiencing. Yet, they are one of the ingredients we use, together with cognition, to recognize our emotions.

The qualities of emotional feelings can be specified along the same dimensions we used to describe the qualities of somatic feelings. First of all, like somatic feelings, emotional feelings are also characterized by a certain intensity which is called *arousal* and is related to the intensity of our bodily changes. This informs us about the significance of the situation. Secondly, they also have a specific valence or *hedonic tone* and can be pleasant or unpleasant and tell us whether the situation we are in is positive or negative for us. Thirdly, emotional feelings can also be more or less short-lived or long lasting. Fourthly, they influence our cognitive processes and bring about specific dispositions for action. Finally and more importantly, as we already mentioned, they tell us whether the object, person or situation that triggers the emotion is good or bad for us (Ortony et al. 1990, chap. 1). As Lazarus puts it, emotions are appraisals of the person-environment relationship and they inform us whether this relationship should be considered beneficial or harmful along various lines (Lazarus 1991, pp. 5–6). Indeed, appraisal is not a unitary evaluation, but is influenced by various factors which together help us evaluate whether our relationship to our environment is beneficial or harmful. Since these feelings are perceived as being caused by/directed toward a specific object, person or situation, they also have a 'localization function', in the sense that they indicate their object (we have emotions toward *something* or *someone*).

Like somatic feelings, emotional feelings can also be considered the product of a monitoring system. The task of this system is to make us aware of emotional states that function as mechanisms for appraisal. In particular, emotional feelings are perceptual signals that convey appraisal information on our relationship with the external environment: whether and how a specific person or situation affects us and our well-being.

As we mentioned before, feelings can never be wrong, since they are merely signals triggered by the outcome of a process: when a process gives a certain output, the corresponding feelings automatically reach our awareness. In spite of this direct link between a process outcome and specific feelings, just like somatic feelings, emotional feelings can be misleading. In fact, it often happens that we have an emotion that is inappropriate for the situation. Think, for example, of cases such us unmotivated (or excessive) anger or fear. It often happens that people feel very scared in reaction to stimuli that should not trigger fear (e.g. in response to a harmless small animal). In these cases, we cannot say that our feelings are mistaken: we actually experience them, and they correctly signal the outcome of an appraisal judgment. For some reason, the stimulus (the small animal) has been interpreted by our cognitive system as extremely dangerous and potentially harmful for us. Extreme fear is the correct reaction to this appraisal judgment. However, from an objective point of view, fear is not the

appropriate emotion for the context and in this sense it is misleading. Especially in the case of complex emotions, emotional feelings can also be misleading in another sense: since they are quite complex and require cognitive elements for interpretation, it is possible that we do not interpret our feelings correctly and we believe, for example, that we feel disrespected, while the feeling we actually experience is jealousy. Even though they can be misleading, emotional feelings are clearly essential for our lives: without them we would not be able to assess our relationship with the external world.

This general overview of the nature and function of somatic and emotional feelings allows us to say something more about epistemic feelings. Like other kinds of feelings epistemic feelings also have an object, i.e. are triggered by something and are felt to be about this something. Specifically, they are triggered by cognitive processes and are felt to be about them: I understand, know or forget something; I am certain or uncertain about something, etc. Their function is to assess these processes or their outcomes from an epistemic point of view and establish, for example, whether a task is easy or difficult, if the various elements are related or unrelated, if we did or did not make a mistake, if we have or do not have certain knowledge, if we are or are not certain of a result, etc.

Like somatic and emotional feelings, epistemic feelings are also possibly conveyed by bodily sensations, even though this aspect is not salient when we analyze their phenomenology. Yet, like other kinds of feelings they are characterized by a specific quality – the way they feel – and it is this quality that carries specific epistemic information. For example, we might have the feeling that we know something, sometimes we even feel it is on the tip of our tongue. It is this specific quality of the sensation we experience that informs us that there is some knowledge we already possess yet cannot retrieve just now. Or we might have the feeling that we have finally understood something: the specific (positive and satisfying) sensation we experience that tells us that all the pieces have finally fallen into place. In this sense, feelings are sources of information and, more importantly, each type of feeling provides a specific type of information by virtue of its specific quality.

The specific qualities of epistemic feelings fall along the same dimensions we identified above for other kinds of feelings. First of all, they are characterized by a certain intensity, which informs us about the importance of our epistemic assessment. Secondly, they have a specific valence or *hedonic tone* that can be pleasant or unpleasant: this tells us whether the epistemic assessment is positive or negative. Thirdly, they can also be more or less short-lived or long lasting. Fourthly, they influence our cognitive processes and bring about specific dispositions for action: the feeling of uncertainty, for example, drives us to look for further evidence; the feeling of difficulty (depending on its intensity) might drive us to pay more attention to a task or give it up; the feeling of error induces us to recheck our work, paying more attention this time, etc. More generally, epistemic feelings orient our thinking by giving us perceptual feedback on cognitive processes. Finally, they are reactive and result from an appraisal of the situation: they provide an assessment of, for example, the accuracy and adequacy of our cognitive processes. As should already be clear from the fact that they are always about something, these feelings also have a 'localization function' in the sense that they indicate the specific cognitive process or the part of this process at stake in every assessment.

Like somatic and emotional feelings, epistemic feelings also might occur as blends of other feelings. The feeling of fluency (whether a process 'flows' easily or not) might be, for example, one of the basic ingredients that contributes to building several more complex feelings: e.g. the feeling of difficulty, the feeling of error and the feeling of certainty (Carruthers 2017). The research usually lists epistemic feelings as if they were all at the same level (as we did here in §2). However, this is probably incorrect. What we need is a taxonomy of epistemic feelings clarifying which are simple or complex and which are components of others. When we have a complex feeling, its quality will be an overall synthesis of all the ingredients. Just as the flavor of a cocktail is not the mere sum of the flavors of the single ingredients you put into it, the quality of complex feelings will not merely consist in the sum of the qualities of the individual components, but result from the blend.

Epistemic feelings are also analogous to somatic and emotional feelings in the sense that although they may be misleading, they are never wrong. Not only are we passive towards our feelings, they are also automatically triggered (or not triggered) by certain cognitive processes. Think for example of familiarity: the feeling that someone or something is familiar to us is triggered by a cognitive mechanism that involves perception and memory; when we recognize that a person, place or object we are perceiving now is identical to a person, place or object we were confronted with in the past, we have a feeling of familiarity. It can happen, for instance, that we perceive someone or something we have already encountered in the past, but we do not experience any feeling of familiarity. What goes wrong in this case is that the recognition process does not result in a positive outcome: we do not recognize the person or the object we are experiencing right now as someone/something we already experienced in the past. Because the output of this process is negative, no feeling of familiarity is triggered. The same applies when in the opposite situation, i.e. when we experience a feeling we shouldn't experience.

We might have the feeling that we know something, but when we actually try to retrieve the content we think we have, we find out that our feeling was misleading. In this case, too, the feeling of knowing occurred as a result of an assessment of our knowledge contents: since we experienced a positive outcome (the content was assessed as available), the feeling was activated. From this point of view, the feelings are only signals that specific cognitive processes resulted in a certain outcome. This outcome immediately triggers the corresponding feeling. However, this outcome might be wrong and thus trigger a feeling that is actually not appropriate. In certain cases – especially when feelings concern complex situations – it is also possible that we misinterpret them: that we experience, for example, a strong feeling of difficulty, but interpret this as a feeling of error. Feelings serve as guides for our cognitive processes; when they are misleading, we are led down the wrong path. Even though epistemic feelings can be misleading, they are still a fundamental guide for our cognitive processes. They allow us to improve ourselves by revising our conclusions or our cognitive processes. Without them we would not have any clue as to the correctness or incorrectness of what we do.

In sum, epistemic feelings, like somatic and emotional feelings, can be considered the product of a monitoring system whose task it is to assess our cognitive processes and contents from various epistemic points of view, e.g. certainty, rightness, clarity,

interest, etc. They are perceptual signals that accomplish a meta-cognitive control function, in the sense that they say something about our thinking process (Proust 2015). Somatic, emotional and epistemic feelings are triggered by different stimuli and result from three different monitoring systems which accomplish divergent evaluative functions: somatic feelings assess states of the body; emotional feelings assess how a situation, an object or a person affects us in various respects; and epistemic feelings assess the epistemic properties of our specific cognitive processes.

4 The Specificity of Epistemic Feelings

This idea that our feelings might derive from multiple systems is not shared by everyone. Indeed, since the output of epistemic monitoring is feelings, some consider epistemic feelings to actually be epistemic emotions (Carruthers 2017). In the following section, we are going to show that the emotional and epistemic monitoring systems must be separate since *it is possible that one works properly while the other does not and in particular, that the epistemic monitoring system works properly, even though the emotional monitoring system does not*. Moreover, we will say something about *what happens to epistemic feelings when the monitoring system that triggers them is impaired*. Finally, we will show that *epistemic feelings guide us to choose specific kinds of reasoning processes over others*: that they determine whether we rely on intuitions rather than on analytic reasoning, that they might not necessarily lead us to make the best choice in a given context, but that they can be trained to provide us with better guidance.

A well-known and widely studied case of emotional dysfunction can be observed in the condition called *alexithymia*. 'Alexithymia' is a neologism derived from old Greek that means lack (a-) of words (lexis) for emotions and moods (thymos). It is generally defined as a subjective inability to describe emotional experiences in words, and it characteristically exhibits four main features: "(1) difficulty identifying and describing subjective feelings; (2) trouble differentiating between feelings and the physical sensations of emotional arousal, (3) limited imaginative processes, and (4) an externally-oriented cognitive style" (Timoney and Holder 2013, p. 1). Even though the symptoms of this condition consist mainly in a linguistic impairment – i.e. people suffering from it cannot verbally describe their emotions – the cause of alexithymia can be traced back to a deficiency of emotion perception. Alexithymia is a condition due to a (more or less severe) lack of emotional awareness: people suffering from it are unaware of their emotional feelings or have difficulties identifying them.

The reason why this condition in relevant to the analysis we have conducted here lies in the fact that the emotional monitoring system of alexithymics does not work properly. In spite of this, people suffering from this condition exhibit a normal capacity to carry out any kind of cognitive task *unrelated to emotions*. Of course, they have a number of deficits with respect to their social life and their mindreading capacities as well as with understanding situations in which emotions play a salient function, but otherwise their cognitive capacities are intact (cf. Krystal 1993). In fact, alexithymics often use some form of reasoning based on external evidence to figure out what

emotions they should experience at a certain moment; this means that they rely on non-emotional reasoning capacities to compensate for their poor ability to identify emotions. As e.g. Kym Maclaren explains, alexithymic individuals "are typically unable to give expression to any emotional experience. They can describe the facts of the situation, and they will often ask what 'one' ought to do in such a situation, but they make no reference to how they, themselves, feel about it […] an alexithymic may reason her way to a conclusion concerning her emotional state, using the physical sensations within her body and the behavior of others as premises for this deduction […] the emotion […] is not something expressed in the first person, but rather hypothesized through a third person process of reasoning" (Maclaren 2006, p. 140).

Alexithymics have trouble identifying their emotions. In spite of this, they are perfectly capable of carrying out cognitive tasks that are unrelated to emotions. Since epistemic feelings are essential for carrying out any cognitive task, their epistemic monitoring system must be working properly. This means that damage to the emotional monitoring system does not mean *per se* that the epistemic monitoring system will also be impaired. If so, emotional and epistemic feelings cannot be one and the same thing. They are both signals brought about by the same vehicle, that is bodily changes, but they are the products of different systems. They arise as the output of a specific monitoring process and their activation depends on the proper functioning of that process.

To substantiate this conclusion, it suffices to show what happens to epistemic feelings when the monitoring system that triggers them is impaired. For this purpose, we can consider the case of another deficit related to alexithymia, i.e. high functioning autism. Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by a number of specific behavioral and cognitive deficits. People suffering from autism exhibit severe issues in social communication and interaction as well as restricted, repetitive patterns of behaviour, interests, or activities (American Psychiatric Association 2013), as well as difficulties in mindreading or empathizing (cf. e.g. Baron-Cohen et al. 1985). Like people suffering from alexithymia, people suffering from autism also have difficulties with emotion processing (in fact, about 50% of autistics exceed the cutoff score for alexithymia and some studies even suggest that the emotional impairment exhibited by autistics is due to alexithymia rather than features of autism per se, cf. e.g. Bird and Cook 2013). But unlike alexithymics who only have an impairment with respect to emotion awareness, autistics also exhibit a particular cognitive style characterized, among other features, by a drive to construct systems – i.e. to identify rules that govern change and allow one to predict how the system will behave (cf. e.g. Baron-Cohen 2006) – and by an impairment of intuitive reasoning which leads to a significant propensity to make use of deliberative reasoning (cf. e.g. De Martino et al. 2008; Brosnan et al. 2017).

Intuitions are rapid and nonconscious thinking processes that work on the basis of heuristics and are independent of working memory and cognitive ability. In the so-called dual-process accounts of human cognition, intuitions are contrasted with a different type of reasoning which is slow, analytic, logical and deliberate (cf. Evans 2008, 2011 for a review). People suffering from autism exhibit enormous difficulties in using this rapid

and intuitive mechanism and compensate for this lack by relying as much as possible on analytic and logical processing.[3] Many explanations of this incapacity to use intuitive reasoning are possible. First of all, the use of intuitions is inconsistent with the systematizing cognitive style of autistics (Brosnan et al. 2014). Secondly, intuitions often involve social and emotional capacities that autistics lack such us mindreading or understanding other people's emotions. Finally and most importantly, people suffering from autism have weak central coherence: in contrast to the general population, they do not try to find coherence amongst the elements they perceive and ignore global information (the big picture) in favor of individual details. If they do a jigsaw puzzle, for example, they do not rely on the picture, but on the forms of the individual pieces. So, for them to do a jigsaw puzzle without a coherent picture would be just as easy. Uta Frith (1989, p. 3) reports the case of a three-year-old autistic child who did jigsaw puzzles upside-down. For the same reason, autistics are much better than neurotypical individuals in finding hidden geometrical figures embedded in pictures: since they do not focus on the overall picture and on its 'significance' (e.g. what we are looking at is a baby carriage), but on the individual pieces that compose the depicted objects, they can easily identify the geometrical figures hidden in them (Frith 1989, p. 152ff). Central coherence – or in more metaphorical terms, the capacity to 'see the big picture' – is essential for intuitive judgments. In fact, intuitions are forms of global processing; flexible heuristics are used to select from all the knowledge available to the subject only that information relevant to reaching a conclusion. All and only the information that is significant for the problem at hand can be taken into consideration, and to make a choice one must have a perspective on the situation and know what is important in that context.

On the other hand, intuitions are crucial for tackling most problems in our everyday life. As, for example, Klin and colleagues (2003, p. 349) suggest, the social world is an "open domain task" that requires an understanding of "a multitude of elements that are more or less important depending on the context of the situation and the person's perceptions, desires, goals and ongoing adjustment. Successful adaptation requires from a person a sense of relative salience of each element in a situation, preferential choices based on priorities learned through experience, and further moment-by-moment adjustments." The neurotypical population usually solves open domain tasks and makes decisions using intuition. Decision making, especially fast decision making, relies on intuition. Unsurprisingly, autistics exhibit great difficulties with (rapid) decision making: they are overwhelmed by information, yet tend to collect too much and have trouble extracting what is relevant; they have problems reaching decisions, and impaired flexibility that leads them to make decisions on the basis of previous choices (Luke et al. 2012; Robic et al. 2015; Vella et al. 2018).

And here we have finally reached our point: previously we argued that epistemic feelings depend on the proper functioning of the process they respond to; this raises the question of what happens to epistemic feelings when the process they respond to is not

[3] As is argued e.g. by Brosnan and colleagues (2016): "Whilst intuitive reasoning is argued to be independent of cognitive abilities, deliberative reasoning is not." In this sense, the success of people suffering from autism who must achieve good results by applying deliberative reasoning strategies will heavily depend on their cognitive abilities; these abilities can be extremely variable even if we consider only the so-called high functioning autistics.

working properly. In autistics, intuitive reasoning processes are impaired: analyzing the feeling that accompanies intuition provides some clues that help answer this question.

First of all, the case of autistics can dispel the idea that, if a cognitive process is impaired or malfunctioning, it does not give rise to any epistemic feelings. In fact, autistics have a lot of feelings in relation to their intuitive reasoning processes. The common trait of the feelings they report is that they are intense and have a negative hedonic tone. Among the sensations most often reported are mental freezing, exhaustion, slowness, discomfort, dislike, higher levels of anxiety, low mood and depression (Lucke et al. 2012; Brosnan et al. 2014; Vella et al. 2018). Anxiety is especially prevalent, likely the result of feelings of uncertainty due to the incapacity to produce intuitive judgments (Boulter et al. 2014).

This negative feedback is certainly also responsible for the fact that autistics try to avoid intuitive reasoning as much as possible and prefer to use deliberate reasoning strategies instead. They also prefer to engage in activities that are more manageable given their cognitive resources. In general, to avoid distress, they prefer to engage in restricted repetitive and stereotyped patterns of behavior, interests and activities: as Boulter and colleagues point out (2014, p. 1398): "knowing all there is to know about a specific restricted interest means that there is little room for unwelcome and uncertain surprises, which may be comforting in a world inherently full of uncertainty." Autistics also prefer tasks that can be solved by using deliberative reasoning strategies, following precise rules or paying attention to specific details. Their cognitive processes support this kind of reasoning quite well, and the steps they go through fluently follow one another. Since they are capable of accomplishing these tasks successfully, they receive reliable, positive and pleasant epistemic feedback through their feelings. Of course, since they have more pleasant feelings when they engage in deliberate reasoning than when they try to solve an issue intuitively, they are always inclined to reason in a slow, analytic and deliberate manner.

This also has some positive collateral effects. Intuitive reasoning processes are useful when we need to solve "open domain tasks" that are too global to be addressed by using an analytic form of reasoning, but they are also prone to errors and biases; this becomes particularly clear when we use our intuition to solve specific logical tasks. In such cases, the reasoning of autistics is less susceptible than that of the neurotypical population to potentially erroneous biases such us the conjunction fallacy (the so-called 'Linda problem') or the assessment of the validity of an argument on the basis of the truth or falsity of the conclusions (Brosnan et al. 2016; De Martino et al. 2008).

Consider the following example:

(i) "All flowers need water. Roses need water. Therefore, roses are flowers" (Brosnan et al. 2017)

People who have never learned the basics of logic or first-time students typically conclude that there is nothing wrong with this kind of reasoning. They jump to this conclusion using an intuitive form of reasoning that relies on real-world experience and focuses on the content of the premises and the conclusion (which is believable) instead of on the process that lead from the first statement to the second (which is invalid).

For the same reason, they tend to consider the following inference as incorrect.

(ii) "Spiders have six legs. Creatures with six legs have wings. Spiders have wings."

In this case, even though the inference is valid from a logical point of view, real-world experience suggests that both the premises and the conclusion are false. Therefore, people who never learned the basics of logic usually conclude that the reasoning process must be wrong. However, since autistics tend to avoid the global form of reasoning based on real-world experience and automatic prefer an analytic, slow and deliberate form of thinking to using intuition, they are less likely than non-autistics to make such mistakes (De Martino et al. 2008; Brosnan et al. 2016; Brosnan et al. 2017).

These observations tell us something interesting about epistemic feelings. Ideally, we should choose the reasoning strategy that is most appropriate to the task we need to accomplish. As Brosnan and colleagues point out: "Typically, intuitive and deliberative reasoning can be applied as appropriate to the perceived demands of the reasoning context." (Brosnan et al. 2016, p. 2122) The key element here is that we *perceive* one kind of reasoning to be more appropriate than another. In fact, epistemic feelings should guide us in the choice of what kind of reasoning process we should activate. However, many factors influence our epistemic feelings. First of all, our epistemic feelings are signals that respond to specific cognitive processes. If one cognitive process works better than another for us (as in the case of deliberate reasoning for autistics), this gives us more reliable and – in general – more pleasant epistemic feelings and we tend to prefer it over others that are less satisfying. So, depending on our cognitive style we will feel better when we apply a specific kind of reasoning process, even if this might not be the appropriate one for the context.

Moreover, if we consider the examples of the inferences reported above, it becomes apparent that the kind of reasoning process that suits non-autistics without logical training is the intuitive approach: they prefer content over process. People who have never learned the basics of logic do not even consider the possibility of checking the procedure instead of the content included in the inferences. However, people with training in logic immediately feel that there is something wrong with the procedure and direct their attention to it, entirely disregarding the content of the premises and the conclusion. These differences in the behavior of people with and without training in a certain field tell us that we have epistemic feelings concerning both the contents and the procedure of our thinking processes. We feel quite certain about the truth/falsity of the specific contents included in these examples and we have also feelings regarding the validity or invalidity of the procedure they follow. If our thinking processes are led by epistemic feelings, then these examples tell us that – in the absence of specific training – our feelings about the content prevail over our feelings about the process. In order to assess reasoning, people rely on their feelings concerning the truth or falsity of the content involved. However, training changes this natural orientation: feelings can be trained and training might help us choose the kind of reasoning most appropriate for a given situation.

5 Concluding Remarks

In this work, we compared somatic, emotional and epistemic feelings and tried to show that they share a number of salient characteristics. However, in spite of their similarities, we described them as the product of three separate systems. The hypothesis we argued for suggests, more specifically, that although somatic, emotional and epistemic feelings all make use of the same kinds of signals (feelings brought about by bodily changes characterized by a phenomenal quality), they are actually triggered by three different automatic monitoring systems.

Somatic feelings are activated by the changes that occur in the body and monitor the effects of the states they signal on our physical health. Emotional feelings are activated by situations that are appraised as relevant for our physical or psychological well-being and monitor the various effects that this situation might have on us. Epistemic feelings are activated by cognitive processes and monitor a number of epistemic features. The feedback provided by feelings is perceptual (non-cognitive) and its qualities (above all its hedonic tone) motivate us to take action: e.g. we react to hunger by seeking for food, we react to fear by freezing or flying, we react to uncertainty by seeking further information.

In the literature, the idea that our feelings might derive from multiple systems is controversial. In particular, epistemic and emotional feelings are sometimes considered to be one and the same thing. Here we tried to show that the emotional and the epistemic monitoring systems must be considered separate because it is possible for one to work properly while the other does not and in particular that the epistemic monitoring system works properly, even though the emotional monitoring system does not. To back up this thesis, we discussed the case of a condition called *alexithymia*, in which the emotional system is impaired and emotional signals are not perceived correctly. Since people suffering from alexithymia do not exhibit any impairment in their reasoning (when this does not involve emotions or the need to understand phenomena related to emotions), we argue that the epistemic feelings of alexithymics work properly.

In the final part of the paper, we consider what happens to epistemic feelings when the process they respond to is not functioning properly. To address this point, we analyze the case of people suffering from autism who are incapable of using intuitive thinking. The case of autistics shows that epistemic feelings always accompany cognitive processes: if a cognitive process is impaired or malfunctioning, the result is intense unpleasant feelings such as mental freezing, exhaustion, slowness, discomfort, dislike, etc. The link between cognitive processes (i.e. types of reasoning) we are not very good at and negative epistemic feelings is then further explored by considering how people familiar and unfamiliar with logic address logical tasks. The kind of reasoning strategy we should adopt (e.g., intuitive vs. rational and deliberate) depends on the task we need to perform. Epistemic feelings should guide us in making this choice. However, if we are more prone to use intuition than to engage in analytic reasoning, we will feel better using intuitive forms of reasoning even in cases in which analytical thinking would be preferable. And yet, feelings can be trained and training (for example, training in logic) might help us choose the kind of reasoning most appropriate for a given situation.

References

American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders, 5th edn. APA, Washington

Arango-Muñoz S (2014) The nature of epistemic feelings. Philos Psychol 27(2):193–211

Baron-Cohen S (2006) The hyper-systemizing, assortative mating theory of autism. Progress Neuro-Psychopharmacol Biol Psychiatry 30:865–872

Baron-Cohen S, Leslie AM, Frith U (1985) Does the autistic child have a "theory of mind"? Cognition 21:37–46

Berthoz A (2000) The brain's sense of movement. Harvard University Press, Cambridge

Bird G, Cook R (2013) Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. Transl Psychiatry 3:e285. https://doi.org/10.1038/tp.2013.61

Boulter C, Freeston M, South M, Rodgers J (2014) Intolerance of uncertainty as a framework for understanding anxiety in children and adolescents with autism spectrum disorders. J Autism Dev Disord 44(6):1391–1402. https://doi.org/10.1007/s10803-013-2001-x

Brosnan M, Chapman E, Ashwin C (2014) Adolescents with autism spectrum disorder show a circumspect reasoning bias rather than 'jumping-to-conclusions'. J Autism Dev Disord 44(3):513–520. https://doi.org/10.1007/s10803-013-1897-5

Brosnan M, Ashwin C, Lewton M (2017) Brief report: intuitive and reflective reasoning in autism spectrum disorder. J Autism Dev Disord 47(8):2595–2601. https://doi.org/10.1007/s10803-017-3131-3

Brosnan M, Lewton M, Ashwin C (2016) Reasoning on the autism spectrum: a dual process theory account. J Autism Dev Disord 46(6):2115–2125. https://doi.org/10.1007/s10803-016-2742-4

Carruthers P (2017) Are epistemic emotions metacognitive? Philos Psychol 30(1–2):58–78

Ceunen E, Vlaeyen JWS, Van Diest I (2016) On the origin of interoception. Front Psychol 7:1–17. https://doi.org/10.3389/fpsyg.2016.00743

Craig AD (2003) Interoception: the sense of the physiological condition of the body. Curr Opin Neurobiol 13:500–505

Craig AD (2009) How do you feel – now? The anterior insula and human awareness. Nat Rev Neurosci 10:59–70

Craig AD (2010) The sentient self. Brain Struct Funct 214:563–577

Damasio A (2010) Self comes to mind. constructing the conscious brain. Vintage Books, New York

De Martino B, Harrison NA, Knafo S, Bird G, Dolan RJ (2008) Explaining enhanced logical consistency during decision making in autism. J Neurosci 28:10746–10750

de Sousa R (2009) Epistemic feelings. Mind Matter 7(2):139–161

Dellantonio S, Pastore L (2017) Internal perception. The role of bodily information in concepts and word mastery. Springer, Heidelberg

Ekman P (1992a) Are there basic emotions? A reply to Ortony and Turner. Psychol Rev 99(3):550–553

Ekman P (1992b) An argument for basic emotions. Cogn Emot 6(3):169–200

Ekman P (1999) Basic emotions. In: Dalgleish T, Power T (eds) The handbook of cognition and emotion. Wiley, New York, pp 45–60

Ekman P (2003) Emotions revealed. Recognizing faces and feelings to improve communication and emotional life. Times Books, New York

Evans JSB (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Ann Rev Psychol 59:255–278

Evans JSB (2011) Dual-process theories of reasoning: contemporary issues and developmental applications. Dev Rev 31:86–102
Frijda NH (1986) The emotions. Cambridge University Press, Cambridge
Frith U (1989) Autism. Blackwell Publishers, US
Klin A, Jones W, Schultz R, Volkmar F (2003) The enactive mind, or from actions to cognition: lessons from autism. Philos Trans R Soc B Biol Sci 358(1430):345–360
Koriat A (2000) The feeling of knowing: some metatheoretical implications for consciousness and control. Conscious Cogn 9(2):149–171. https://doi.org/10.1006/ccog.2000.0433
Krystal H (1993) Integration and self-healing. affect, trauma, alexithymia. The Analytic Press, Hillsdale
Lane RD, Schwartz GE (1987) Levels of emotional awareness: a cognitive-developmental theory and its application to psychopathology. Am J Psychiat 144:133–143
Lazarus RS (1991) Emotion and adaptation. Oxford University Press, New York
Luke L, Clare IC, Ring H, Redley M, Watson P (2012) Decision-making difficulties experienced by adults with autism spectrum conditions. Autism 16(6):612–621. https://doi.org/10.1177/1362361311415876
Maclaren K (2006) Emotional disorder and the mind-body problem: a case study of alexithymia. Chiasmi Int. 8:139–155
Michaelian K, Arango-Muñoz S (2014) Epistemic feelings and epistemic emotions. Philos Inquiries 2(1):97–122
Oatley K, Johnson-Laird PN (1987) Towards a cognitive theory of emotion. Cogn Emot 1:20–50
Ortony A, Clore GL, Collins A (1990) The cognitive structure of emotions. Cambridge University Press, Cambrigdge
Prinz JJ (2004) Gut reactions. A perceptual theory of emotions. MIT Press, Cambridge
Proust J (2015) The representational structure of feelings. In: Metzinger T, Windt JM (eds) Open MIND: 31(T). Frankfurt am Main: Mind
Robic S, Sonié S, Fonlupt P, Henaff MA, Touil N, Coricelli G, Mattout J, Schmitz C (2015) Decision-making in a changing world: a study in autism spectrum disorders. J Autism Dev Disord 45(6):1603–1613. https://doi.org/10.1007/s10803-014-2311-7
Scherer KR, Schorr A, Johnstone T (2001) Appraisal processes in emotion: theory, methods, research. Oxford University, Oxford
Schwarz N (2010) Feelings-as-information theory. In: Van Lange PAM, Kruglanski AW, Higgins ET (eds) Handbook of theories of social psychology, vol. 1. https://doi.org/10.4135/9781446249215.n15
Timoney LR, Holder MD (2013) Emotional processing: deficits and happiness. Assessing the measurement, correlates, and well-being of people with Alexithymia. Springer, Heidelberg
Vella L, Ring HA, Aitken MR, Watson PC, Presland A, Clare IC (2018) Understanding self-reported difficulties in decision-making by people with autism spectrum disorders. Autism 22(5):549–559. https://doi.org/10.1177/1362361316687988

A Model for the Interlock Between Propositional and Motor Formats

Gabriele Ferretti(✉) and Silvano Zipoli Caiani

Department of Humanities, University of Florence, Florence, Italy
fairg@live.it

Abstract. One of the most important tasks of philosophy of mind is the investigation of the nature of our mental states. However, mental states come in different formats. In this respect, one of the most interesting problems in contemporary philosophy of mind is determining how mental states coming in different formats can interlock. Such a problem has generated two parallel debates, especially when we try to describe the nature of practical knowledge in skilled motor action: the one about *Intellectualism* and the one about the *Interface Problem*. Both of these two debates, which are at the crossroad between philosophy of mind and philosophy of action, seem to share a common problem. Action performance requires the interplay between two different representational formats, the practical and the propositional. If so, how can we account for such an interlock between these two different formats? We mention these two debates as starting point, in order to highlight the importance of providing an account capable of explaining the relation between different mental states, especially in the case of propositional and pragmatic, motor states. However, we do not want to directly tackle these two literatures here. Rather, we want to sketch a possible solution for such a problem, which can be beneficial for both the literatures. We suggest that motor and propositional states interlock through the *motor mediation* of *action concepts*. Our account is cashed out from the philosophical analysis of empirical evidence showing that the information processing that is responsible for the generation of the representations of action concepts is strictly related to the motor processing that is responsible for generating the appropriate representations recruited in the planning and execution of motor behaviors.

1 Introduction

One of the most interesting tasks of philosophy of mind is the investigation of the nature of mental states. But, mental states come in different formats. I can visualize in my visual imagery a red apple, I can think in propositional-like representations where I would like to spend my holydays, I can remember what it is like to be thirsty, I can feel how hard it is to lift a weight. In this respect, one of the most interesting issues in contemporary philosophy of mind is determining how mental states coming in different formats can interlock. Notably, the debate is nowadays about how *pragmatic*, *motor* formats and *propositional* formats can interlock. This is very interesting, as this question has generated two parallel debates, especially concerning the nature of practical knowledge in skilled action.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 427–440, 2019.
https://doi.org/10.1007/978-3-030-32722-4_24

This first debate is the one generated by Ryle (1949), who has firstly suggested that practical knowledge captured by the sentence "Mark knows how to run" is a different knowledge with respect to the propositional knowledge captured by the sentence "James knows that Rome is beautiful". This posits a basic difference between practical and propositional knowledge. In the last years, however, supporters of *intellectualism* have suggested that sentences of the first kind capture nothing but propositional states of the second kind. Saying that "Alfred knows *how* to swim" is nothing but saying that, for a given way *w* to swim, "Alfred knows *that w*" is a way to swim Stanley and Williamson (2001).

Such a knowledge that *w* is a way to swim comes, however, under a *pragmatic mode of presentation* (Stanley 2011). Anti-*intellectualists* (Fridland 2013, 2016) are today trying to dismount such an account, by suggesting that, since practical knowledge requires special mental non-propositional states such motor representations (henceforth: MRs) (Jeannerod 2006; Nanay 2013; Ferretti 2016c, b, d, 2019, Brozzo 2017), the fact that "Michael knows *how* to do a triple gainer" does not amount to posses any knowledge *that* a given way *w* is a way to do a triple gainer. Practical knowledge requires non-propositional states like MRs simply because, on the one hand, propositional knowledge would lead to infinite regress (Fridland 2013) and, on the other, propositional states cannot describe, due to their formats, the enormous kinematic and biomechanical complexity lying behind practical knowledge. Indeed, doing a triple gainer requires the possession of adequate MRs related to the realization of a triple gainer (Butterfill and Sinigaglia 2014; Sinigaglia and Butterfill 2015; Davidson 1963). We do not want to offer an argument in support of one of these positions here. However, the reader should note that, independently of whether we embrace anti-*intellectualist* at the expense of *intellectualism*, we need to explain how propositional and practical knowledge can be related. If, on the one hand, we embrace *intellectualism*, it is not clear how a propositional state, thought coming under a pragmatic mode of presentation, can generate appropriate motor commands. On the other, if we embrace anti-*intellectualism*, we need to explain why, though a skill is guided by a practical knowledge, we can connect such a practical knowledge, for example of how to swim, to our propositional states about the fact that *w* is a way to swim. A final account concerning these two explanations are not offered in the literature.

This is even clearer in the light of another debate on a similar issue. According to a classical view in action theory, appropriate action performance requires the agent to link appropriate desires with appropriate beliefs, this is the famous belief-desire model (Bratman 1999; Davidson 1963). Recently, however, more empirical informed accounts in the literature on action theory have suggested that, as we saw also above, propositional states cannot capture the complexity of motor behavior, and that such a complexity can be captured only by non-propositional mental states (Nanay 2013). So, when studying intentional action, we no long need to understand how a given desire is related to a given belief. Rather, we should explain how desires and *intentions*, relate to MRs. This is, however, a big problem, as intentions are mental phenomena coming in a propositional format *par* excellence. Intentions are usually conceived as having a *propositional, sentence-like format* (Butterfill and Sinigaglia 2014: 130; Bratman 1987; Mele 1992). For this reason, they can be featured as premises or consequences of a

piece of practical reasoning. However, MRs are not, they are provided with a motor format (Jeannerod 2006; Butterfill and Sinigaglia 2014). This is for a simple reason: MRs allow us to properly represent all the motor aspects of a specific action in a given motor situation. Only the motor format allows providing information concerning all the visuomotor, biomechanical and kinematic aspects of action needed in order to obtain the intended motor performance (Butterfill and Sinigaglia 2014: 130; Jeannerod 2006; Jacob and Jeannerod 2003; Ferretti 2016c; Zipoli Caiani and Ferretti 2017; Fridland 2016)[1]. In this respect, the propositional format is not suitable to deliver information with a so high degree of specificity, as required in order for motor action to properly unfold (Butterfill and Sinigaglia 2014; Shepherd 2017; Burnston 2017). However, standard views assume that an adequate account of the purposiveness of actions needs coordination between intentions and MRs, (Bach 1978; Searle 1983; Mele 1992; Pacherie 2000; Shepherd 2017: Sect. 1; Burnston 2017: Sect. 2). However, such an explanation has to face a crucial issue: since MRs do not have a propositional format, it is not clear how they can interlock with propositional intentions. This issue is at the core of the *interface problem*, and has been recently addressed by prominent authors in the field of philosophy of mind (Butterfill and Sinigaglia, 2014; Mylopoulos and Pacherie 2016; Shepherd 2017; Burnston 2017). A final explanation for such an interlock is not offered in the literature.

Both of these two debates, which are at the crossroad between philosophy of mind and philosophy of action, seem to share a common problem, which concerns how we can account for the interlock between practical formats, especially those at the basis of our MRs, and propositional formats, at the basis of our propositional knowledge, as well as of our intentions. Now, we introduced these two debates in order to highlight the importance of finding an account able to explain the relation between propositional and pragmatic, motor states. However, we do not want to directly tackle these two literatures. Rather, we want to sketch a possible solution for such a problem that can be beneficial for both literatures. We suggest that motor and propositional states interlock through the *motor mediation* of *action concepts*. The basic idea is that action concepts are constituents of a propositional structure, like an intention, but have the same motor format of a MR, as with it they share the same neural realizer, whose processing comes in a motor format. Our account can be useful for the *intellectualist*, who still needs to explain how a propositional state, thought coming under a pragmatic mode of presentation, can generate appropriate motor commands (Fridland 2013), but also to the anti-*intellectualist*, who still needs to explain why, though a skill is guided by practical knowledge, we can connect such a practical knowledge, for example of how to swim, to our propositional states about the fact that *w* is a way to swim. Finally, our account is beneficial to those interested in the *interface problem*, as there seems to be no final solution to it (Ferretti and Zipoli Caiani 2018).

[1] For a philosophical review of the several computational aspects of MRs see (Ferretti 2016a, b, c, d, 2017, 2019, forthcoming; Ferretti and Zipoli Caiani 2018; Ferretti and Chinellato 2019; Chinellato et al. 2019; Zipoli et al. 2017).

Our account of such an interlock between these different formats is cashed out from the philosophical analysis of experimental evidence showing that the information processing that is responsible for the generation of the representations of action concepts is strictly related to the motor processing that is responsible for generating the appropriate representations recruited in the correct planning of motor behaviors. The philosophical analysis of this evidence will allow us to explain how motor and propositional states can interlock.

2 Propositional and Motor Formats at the Crossroad: The Evidence from Neuroscience of Action and Language

Here we provide an account of how propositional and pragmatic formats can interlock. The basic idea is that propositional states and pragmatic-motoric states can interact through the motor mediation of action concepts: executable action concepts that constitute an intention, and MRs, share similar neural correlates and, arguably, the same representational format. In turn, this argument will be the basis for the claim that propositional states, such as intentions, and MRs interlock through action concepts.

First off, it should be noted that the idea that cognition relies on the sensorimotor system is already popular under the name "grounded cognition" (Barsalou 2008). The notion of grounded cognition suggests that cognitive representations are not *amodal* vehicles that exist independently of the brain's sensory and motor systems, but are deeply grounded in the agent's perceptual and action systems (Barsalou 1999; Glenberg and Kaschak 2002; Decety and Grèzes 2006). This means that the concepts corresponding to motor actions and to action-related objects may be represented in areas of the brain specialized for planning and executing motor actions.

This latter aspect is of prominent relevance for the point at stake here. It is generally agreed, indeed, that the possession of action concepts is associated with the ability to inferentially proceed from action planning to action execution (see, for example, Bratman 1999; Pacherie 2008; Searle 1983). Accordingly, action concepts figure as *mediators* between propositional states, such as the agent's beliefs and desires underlying intentional action plans, and the MRs that are immediately antecedent to action execution. However, as already illustrated (§1), here is where the problems of interface problem pop out.

In light of this, there is an increasing interest in understanding the nature of the action concepts mediating between propositional intentions and motor representations (e.g., Fridland 2016; Levy 2015; Zipoli Caiani and Ferretti 2017). To solve these problems related to the interface between different formats, we propose that, while intentions are built in a propositional format, they have, as proper constituents action concepts, which are built in a non propositional format. Since action concepts are shaped in a motor format, they can serve as a bridge between high-level cognition and motor behavior. Now, the challenge here is to show that the representation of action concepts has, *contra* the classical view, a pragmatic motor format. Note also that it has been commonly suggested that, usually, when a representation not directly involved in overt execution is generated by the activation of the motor system, this is reliably for

the fact that it is recruiting a cognitive process coming in a motor format (Jeannerod 2006: Sect. 2). This paper supports this view when talking about action concepts.

Now, in order to support our view that action concepts have the same pragmatic format of MRs, we provide, *step by step*, three types of evidence. First, we show that the representation of action concepts and the representation of motor outcomes, performed by MRs, share the same substrate. Second, we argue that action concepts are grounded in the sensorimotor system depending on the mode of execution that is proper for the type of action to which they relate. This suggests how the semantic variability of verbal instructions (e.g., grasping the glass or grasping the pen) is made possible by the different action purposes that are encoded by the intention. Third, we argue that the inhibition or the arousal of somatotopic areas of the motor system induces consequences not only for action execution but also for the processing of action concepts. The following sections will be devoted to discussing such evidence.

2.1 From Action Concepts to Action Execution and Back

It is almost twenty years that the contribution of the sensorimotor system to the processing of action concepts has been under the spotlight of experimental research. Recently, an increasing number of results have been published showing that the motor execution of an action can be modulated by categorization tasks concerning the use of related action verbs. The importance of this evidence relies on the fact that it provides support to the hypothesis that action concept understanding is grounded in the functioning of the sensorimotor system by means of their common representational format.

Among the numerous strategies adopted to support this idea, there are those concerning the relation between language understanding and action execution. A piece of evidence of this sort is provided by Glover et al. (2004). Here, scholars employed words indirectly related to the size of the target object, such as ‘apple’ for a prototypically large object, or ‘grape’ as a prototypically small object. The authors observed that the induced categorization of the target as a paradigmatically large object led the agent to adopt a larger grip aperture than that used when the agent was induced to categorize the target as a small object (see also Glover and Dixon 2002).

A more direct evidence that the processing of action concepts modulates action execution is offered by Boulenger et al. (2006). In this case, subjects were asked to perform a reaching action concurrently or successively to lexical decision tasks concerning action verbs (e.g., to paint, to jump, and to cry) or nouns of concrete entities that cannot be manipulated (e.g., star, cliff, and meadow). The analyses of the movement parameters revealed that within 200 ms after onset, wrist acceleration peaks appeared significantly later and were smaller following the display of action verbs than during displays of the nouns of non-manipulable objects. Thus, since a wrist acceleration peak is indicative of initial muscular contractions, the measurement of longer latency and smaller amplitude suggests that processing action concepts interferes with the execution of the movement itself (see also Nazir et al. 2008). More recently, Andres et al. (2015) provided evidence of the verb-effector compatibility effect, showing the interference between action concept processing and the use of compatible effectors. Moreover, Klepp et al. (2017) have shown that the semantic processing of high

effector-specific movement verbs in German language, such as *rubbeln* (to rub) or *springen* (to jump), induces a body-parts specific facilitation effect.

Additionally, there is also behavioral evidence that clearly supports the view that the processing of action concepts can be shaped by the activity of the motor apparatus. For example, Fernandino et al. (2013a) provided evidence that the disorder of the motor system that characterizes patients suffering from Parkinson Disease is associated with specific impairments in the processing of action-related sentences. They showed that Parkinson Disease patients exhibit significantly longer reaction times when processing action-related sentences that contain action concepts, if compared to control subjects, providing support to the claim that the motor system plays a functional role in the processing of action concepts (see also Fernandino et al. 2013b; Desai et al. 2015; Bidet-Ildei 2017).[2]

The evidence previously reviewed supports the hypothesis that processing action concepts shapes action preparation and execution. Nevertheless, since we assume that action concepts and motor representation share the same representational resources, we expect that a priming effect also occurs in the reverse direction, meaning that the execution of motor actions should interfere with the processing of the related action concepts.

For example, in a classic study by Lindemann et al. (2006), participants were asked to make lexical decisions concerning action-related words or pseudo-words in a go/no-go task paradigm (i.e., valid word 'go' and pseudo-word 'no-go') after having prepared for a specific action that they execute only after word presentation. The results show that the response latencies were reduced if the action-related concepts expressed by the words were consistent with the previously prepared action (see also Rueschemeyer et al. 2010). Moreover, Van Elk et al. (2008) investigated the influence of motor preparation on the time course of action-related concepts comprehension, showing the presence of a facilitation effect (see also van Dam et al. 2012).

Since the sensorimotor system modulates the understanding of action related sentences, it is conceivable that the plasticity of the motor system affects the processing of concepts. According to this hypothesis, the acquisition of novel motor behaviors should improve the agent's ability to recognize and classify different types of actions. For example, Locatelli et al. (2012) trained participants to learn new manual actions and measured their performance on a semantic judgment task after and before the motor training. The results showed that the reaction times significantly decreased after the motor training, indicating an improvement of the comprehension performance (see also Glenberg et al. 2008).

Interestingly, the fact that the acquisition of action-related categorization abilities can be modulated by the improvement of the related action skills can be accounted for

[2] The best explanation of the above-mentioned behavioral evidence is that the processing of action concepts relies on the functioning of the motor system at the neural level, and therefore that action concepts are motorically formatted. This is supported by the following two facts: (1) the priming effect is usually measured within 200 ms after the stimulus onset. This evidence is not compatible with the hypothesis that the agent is imagining the action or that he/she is performing some instance of indirect inferential processing; (2) our interpretation is compatible with the evidence that action concepts processing somatotopically and functionally activates the motor system.

by means of the assumption that action categories and motor schemas share the same representational substrate. Indeed, if executable action concepts and action schemas were encoded by different representational occurrences, there would be no reason to expect a correlation between action enhancement and semantic competences.

To sum up, the previously introduced set of behavioral evidence clearly supports the view that the semantic information of action concepts and the pragmatic information involved in action execution are mapped together, namely, on the same representational substrate. Now, the following sections provide evidence that action concepts processing and MRs not only share the same neural resources, but also that they are *somatotopically* and *functionally* related.

2.2 Action Concepts and the Motor System

In the previous section, we reported behavioral evidence showing that processing action concepts and executing motor actions are not two reciprocally independent cognitive tasks. Our hypothesis is that this mutual influence can be read as evidence that the representations of action concepts and action outcomes share a common cortical substrate. Indeed, showing that processing action concepts involve the activation of the motor system provides evidence for the fact that the former may involve the same motor information delivered by the latter.

The relationship between concept processing and sensorimotor processing is a central topic in cognitive neuroscience, here, several studies investigated the involvement of the motor system during the comprehension of verbs, phrases, or sentences concerning motor actions. Importantly, the somatotopic organization of the human motor cortical system has permitted to test the above hypotheses, providing evidence that action concept representations overlap with the neural substrates involved in performing an action (Coello and Fischer 2015).

Interestingly, empirical investigations have shown a systematic correlation between the processing of action concepts and the functioning of motor areas semantically related to such concepts (Coello and Fischer 2015; Desai et al. 2010; Desai et al. 2013; Kemmerer et al. 2008). At the core of this view is the evidence that the activation of the motor system during action verbs processing is somatotopically organized, so that the processing of action concepts functionally activates the motor areas involved in the execution of the related actions.

Famously, Hauk et al. (2004) found that passive reading of hand, foot, and mouth action verbs (e.g., to pick, to kick, and to lick) activates ventral face/mouth, lateral arm/hand, and dorsal leg/foot motor regions, respectively (see also, Boulenger et al. 2009, 2012; van Elk et al. 2010). While Tettamanti et al. (2005) showed that the same somatotopic organization is also preserved in sentence processing. Moreover, Fargier et al. (2012) found that subjects who have learned new action-related concepts show a neurophysiological sign in the motor cortex activity (precisely for the processing of these words after learning.

Furthermore, Wu et al. (2013) showed that this somatotopic organization of the representation of action concepts concerns the common way agents use and understand action categories in different languages and cultures. of action concepts have also been reported for sentences and verbs, in French, Italian, German, and Finnish (Pulvermüller

2013). Recently, in accordance with previous behavioral evidence, somatotopic activation in the sensorimotor system has also been found for concepts related to objects that afford hand and mouth movements, such as "fork", which activates hand motor regions, and food words, such as "bread", activating face motor regions (Carota et al. 2012).

In accordance with the behavioral evidence previously introduced (e.g., Casile and Giese 2006; Locatelli et al. 2012), additional studies have shown that new motor experience can facilitate action concept understanding by enhancing the activity of the brain motor regions in intentional planning. For example, Beilock et al. (2008) showed that subjects with hockey experience were facilitated in the comprehension of hockey sentences by increasing the activity in the left dorsal premotor cortex and decreasing the activity in bilateral sensorimotor cortex. Similarly, Tomasino et al. (2013) found that the somatotopic activity within the sensorimotor areas induced by action concept processing was a function of the action feasibility and agent's expertise.

This section has shown that the cognitive processing of action concepts is grounded in the activity of the sensorimotor cortex. Accordingly, we have seen that the semantic variability that characterize the processing of action concepts (e.g., *grasping* the glass or *grasping* the pen) is made possible by the different purposes that are encoded by the agent's intention.

2.3 Action Concepts and the Functional Encoding of the Motoric

We have already shown that action concepts are somatotopically represented within the agent's motor system. However, the mere activation of a cortical area does not guarantee that this area is functionally involved in the generation of action performance. It is, therefore, particularly important to confirm that the functioning of the motor system has actual relevance in action concept processing (Leshinskaya and Caramazza 2014; Mahon and Caramazza 2008).

Using magnetic stimulation technique (TMS), Buccino et al. (2005) showed that motor evoked potentials (MEPs) recorded from hand muscles were specifically modulated by listening to sentences concerning hand-related actions, whereas MEPs recorded from foot muscles were modulated by listening to sentences concerning foot-related actions. More recently, Innocenti et al. (2014) verified an analogous result, showing that verbs describing actions rather than concepts modulated the primary motor cortex excitability.

Although we have clearly shown that motor areas are activated when people process action concepts, we still need to establish whether such activation is necessary for concept understanding.

In this regard, the study of brain lesions of the motor system is of particular relevance. For example, Ibáñez et al. (2013) showed that the processing of action concepts in Parkinson's disease patients requires a preserved functionality of the agent's motor repertoire. Differently, Desai et al. (2015) tested stroke patients founding that that the degree of impairment in reaching performance, due to a lesion in the hand/arm motor areas, allows to predict the degree of selective impairment for processing action words concerning reaching actions. Moreover, Kemmerer et al. (2012) explored the behavioral patterns of several brain-damaged patients, founding a

correlation between the presence of lesions on somatotopic related areas of the motor system and the inability to process action concepts (see also Bak and Chandran 2012).

To sum up, this set of evidence suggests that the somatotopic activity of the motor system is not a mere consequence of action concepts processing, but rather a *constitutive* part of our conceptual representation of executable actions. Indeed, if the recruitment of motor cortical resources were only a subsequent step of the processing of the executable action concepts involved in our intentions, there would be no reason to hypothesize that an impairment of the agent's motor repertoire also affects the ability to use and understand executable action concepts. But since the representation of our executable action concepts is mapped directly on the representation of executable actions, the impairment in the processing of the latter involves the impairment in the processing of the former.

At this point, we have reported all the sets of evidence in support of our argument: the representation of action concepts has the same representational format of MRs: a motor format. But executable action concepts are at the basis of our propositional reasoning about action performance and pragmatic abilities (Mylopoulos and Pacherie 2016; Butterfill and Sinigaglia 2014; Mylopoulos and Pacherie 2016; Burnston 2017; Shepherd 2017). If so, the special motoric relation between action concepts and MRs also explains how propositional and motor states can interlock: they can do so by means of the *motor bridge* provided by an action concept. When I intend to perform a given action, and I propositionally reason on that action, these two propositional have at their basis an action concept, i.e. I intend to *grasp* the bottle and I know I can perform a power grip. These propositional states interlock with the corresponding MR of grasping because the verb *to grasp* recruits the same motor resources, within the motor system, which also are the vehicles of the MR. For this reason, there is usually a *motor harmony* between our propositional reasoning about a given action and our overt execution of that same action.

3 Conclusion

One of the most intriguing puzzles in the current philosophy of mind is to understand how the different representational formats involved in cognition come together, giving rise to experience and intentional behavior. In this paper we addressed two particular versions of this issue in the literature. A particular version of this issue, also known as *the interface problem*, concerns the problem of understanding how intentions, that come in a propositional format, hook up with MRs, that come in a motoric-pragmatic format. To solve the interface problem, we proposed that propositional states intentions and motor representations interlock through action concepts. This view suggests the relevance of the executable action concepts, both for the constitution of propositional states, like intentions, and for the prescription of a MRs. Notably, this view states that the representational structure of executable action concepts shares the same representational format that pertains to MRs. This amounts to saying that the action concepts involved in our propositional states, like intentions, and the MRs that allow the execution of an action, share the same non-propositional motor format used both when we plan actions and during covert or overt action performance. In other words,

propositional and pragmatic interface is possible because "action concepts are such stuff as MRs are made on". This 'stuff' is the plethora of the motoric resources of our cognitive system.

Our thesis has been supported, step by step, by a cluster of evidence showing that that action concepts and MRs share the same neural correlates, and arguably the same motor format. Indeed, the recruitment of somatotopic areas of the motor cortex is crucially involved both when we think in motor terms with action concepts and when we plan or perform actions through motor representations. Thus, since the activity of these neural correlates gives raise to representations that are usually considered as built in a motor format, we concluded that executable action concepts are the *motor mediators* between intentions and MRs.

It should also be noted that our view shows relevant differences compared with other relevant proposals in the debate. In particular, unlike the views by Butterfill and Sinigaglia (2014), Mylopoulos and Pacherie (2016), Burnston (2017) and Shepherd (2017), our view is able to really suggest why we do not need a translation process between propositional and non-propositional formats – something not explained by the other accounts. Moreover, our proposal is well supported by empirical evidence at different levels of analysis and is also much simpler and more complete than the other offers on the market.

Furthermore, as announced at the beginning, our account can also be very useful for the debate on *intellectualism*, which is another issue in the literature, and which aims to explain whether and how pragmatic knowledge can be reduced to propositional knowledge. In particular, if we embrace *intellectualism*, namely the view that knowing-how really is a form of knowing-that, we still need to explain how a propositional state, thought coming under a pragmatic mode of presentation, can generate appropriate motor commands (for a review on this point see Fridland 2013). But also if we embrace anti-*intellectualism*, namely the view that knowing-how cannot bye reduced to knowing-that at all, we still need to explain how such a practical knowledge about a given action performance can march in step with the propositional states we can form about that action performance. Our account allows the scholar interested in such a debate to understand how knowing-how subserved by motor states and propositional knowledge can interlock.

Summing up, our account is very beneficial for the literature on the format of representations. First of all, its outcome is crucial for two debates in the spotlight of philosophy of mind: the one on *intellectualism* and the one on the *interface problem*. Second, it paves the way for an understanding the relation between pragmatic and propositional formats, independently of which is the specific position we want to embrace in each of these two literatures[3].

[3] The authors are listed in alphabetical order and contributed equally to the article. We wish to thank those scholars who discussed with us about these topics: Bence Nanay, Chiara Brozzo, Joshua Shepherd, Carlotta Pavese, Andrea Borghini, Brian B. Glenney. We thank the audience at the Conference of the Italian Association for Cognitive Science (AISC) in December 2017 for the very insightful comments on this topic. We also thank two anonymous reviewers for the comments.

References

Andres M, Finocchiaro C, Buiatti M, Piazza M (2015) Contribution of motor representations to action verb processing. Cognition 134:174–184

Bach K (1978) A representational theory of action. Philos Stud 34:361–379

Bak TH, Chandran S (2012) What wires together dies together: verbs, actions and neurodegeneration in motor neuron disease. Cortex 48(7):936–944. https://doi.org/10.1016/j.cortex.2011.07.008

Barsalou LW (1999) Perceptual symbol systems. Behav Brain Sci 22(4):577-609-660

Barsalou LW (2008) Grounded cognition. Annu Rev Psychol 59(1):617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Beilock SL, Lyons IM, Mattarella-Micke A, Nusbaum HC, Small SL (2008) Sports experience changes the neural processing of action language. Proc Nat Acad Sci 105(36):13269–13273. https://doi.org/10.1073/pnas.0803424105

Bidet-Ildei C, Meugnot A, Beauprez SA, Gimenes M, Toussaint L (2017) Short-term upper limb immobilization affects action-word understanding. J Exp Psychol Learn Mem Cogn 43 (7):1129–1139

Boulenger V, Roy AC, Paulignan Y, Deprez V, Jeannerod M, Nazir TA (2006) Cross-talk between language processes and overt motor behavior in the first 200 msec of processing. J Cogn Neurosci 18(10):1607–1615. https://doi.org/10.1162/jocn.2006.18.10.1607

Bratman M (1987) Intention, plans, and practical reason. Harvard University Press

Bratman ME (1999) Intention, Plans, and Practical Reason. Center for the Study of Language and Information, Stanford

Buccino G, Riggio L, Melli G, Binkofski F, Gallese V, Rizzolatti G (2005) Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. Cogn Brain Res 24(3):355–363. https://doi.org/10.1016/j.cogbrainres.2005.02.020

Brozzo C (2017) Motor intentions: how intentions and motor representations come together. Mind Lang 32(2):231–256

Burnston DC (2017) Interface problems in the explanation of action. Philos Explor 20(2):242–258

Butterfill SA, Sinigaglia C (2014) Intention and motor representation in purposive action: intention and motor representation in purposive action. Philos Phenomenol Res 88(1):119–145. https://doi.org/10.1111/j.1933-1592.2012.00604.x

Campbell J (1994) Past, space and self. MIT Press, Cambridge

Carota F, Moseley R, Pulvermüller F (2012) Body-part-specific representations of semantic noun categories. J Cogn Neurosci 24(6):1492–1509. https://doi.org/10.1162/jocn_a_00219

Casile A, Giese MA (2006) Nonvisual motor training influences biological motion perception. Curr Biol 16(1):69–74. https://doi.org/10.1016/j.cub.2005.10.071

Chinellato E, Ferretti G, Irving L (2019) Affective visuomotor interaction: a functional model for socially competent robot grasping. In: Martinez-Hernandez U et al (eds) Biomimetic and biohybrid systems, vol 11556. Living machines 2019. Lecture notes in computer science. Springer, Cham

Coello Y, Fischer MH (2015) Perceptual and emotional embodiment: foundations of embodied cognition. Routledge, Abingdon

Decety J, Grèzes J (2006) The power of simulation: imagining one's own and other's behavior. Brain Res 1079(1):4–14. https://doi.org/10.1016/j.brainres.2005.12.115

Desai RH, Binder JR, Conant LL, Seidenberg MS (2010) Activation of sensory-motor areas in sentence comprehension. Cereb Cortex 20(2):468–478. https://doi.org/10.1093/cercor/bhp115

Desai RH, Conant LL, Binder JR, Park H, Seidenberg MS (2013) A piece of the action: modulation of sensory-motor regions by action idioms and metaphors. NeuroImage 83:862–869. https://doi.org/10.1016/j.neuroimage.2013.07.044

Desai RH, Herter T, Riccardi N, Rorden C, Fridriksson J (2015) Concepts within reach: action performance predicts action language processing in stroke. Neuropsychologia 71:217–224. https://doi.org/10.1016/j.neuropsychologia.2015.04.006

Davidson D (1963) Actions, reasons and causes. J Philos 60:685–700

Fargier R, Paulignan Y, Boulenger V, Monaghan P, Reboul A, Nazir TA (2012) Learning to associate novel words with motor actions: language-induced motor activity following short training. Cortex 48(7):888–899. https://doi.org/10.1016/j.cortex.2011.07.003

Ferretti G (2019) Visual phenomenology versus visuomotor imagery: how can we be aware of action properties? Synthese. https://doi.org/10.1007/s11229-019-02282-x

Ferretti G, Zipoli Caiani S (2018) Solving the interface problem without translation: the same format thesis. Pac Philos Q. https://doi.org/10.1111/papq.12243

Ferretti G (2017) Two visual systems in molyneux subjects. Phenomenol Cogn Sci 17(4):643–679. https://doi.org/10.1007/s11097-017-9533-z

Ferretti G (forthcoming) Why trompe l'oeils deceive our visual experience. J Aesthet Art Crit

Ferretti G (2016a) Visual feeling of presence. Pac Philos Q. https://doi.org/10.1111/papq.12170

Ferretti G (2016b) Neurophysiological states and perceptual representations: the case of action properties detected by the ventro-dorsal visual stream. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology, vol 27. Studies in applied philosophy, epistemology and rational ethics. Springer, Cham, pp 179–203

Ferretti G (2016c) Pictures, action properties and motor related effects. Synth Spec Issue Neurosci Philos 193(12):3787–3817. https://doi.org/10.1007/s11229-016-1097-x

Ferretti G (2016d) Through the forest of motor representations. Conscious Cogn 43:177–196

Ferretti G, Chinellato E (2019) Can our robots rely on an emotionally charged vision-for-action? An embodied model for neurorobotics. In: Vallverdú J, Müller V (eds) Blended cognition, the robotic challenge, vol 12. Springer series in cognitive and neural systems. Springer, Cham

Fernandino L, Conant LL, Binder JR, Blindauer K, Hiner B, Spangler K, Desai RH (2013a) Where is the action? Action sentence processing in Parkinson's disease. Neuropsychologia 51 (8):1510–1517

Fernandino L, Conant LL, Binder JR, Blindauer K, Hiner B, Spangler K, Desai RH (2013b) Parkinson's disease disrupts both automatic and controlled processing of action verbs. Brain Lang 127(1):65–74

Fridland E (2013) Problems with intellectualism. Philos Stud 165(3):879–891. https://doi.org/10.1007/s11098-012-9994-4

Fridland E (2016) Skill and motor control: intelligence all the way down. Philos Stud. https://doi.org/10.1007/s11098-016-0771-7

Glenberg AM, Kaschak MP (2002) Grounding language in action. Psychon Bull Rev 9(3):558–565

Glenberg AM, Sato M, Cattaneo L (2008) Use-induced motor plasticity affects the processing of abstract and concrete language. Curr Biol 18(7):R290–R291. https://doi.org/10.1016/j.cub.2008.02.036

Glover S, Dixon P (2002) Semantics affect the planning but not control of grasping. Exp Brain Res 146(3):383–387. https://doi.org/10.1007/s00221-002-1222-6

Glover S, Rosenbaum DA, Graham J, Dixon P (2004) Grasping the meaning of words. Exp Brain Res 154(1):103–108. https://doi.org/10.1007/s00221-003-1659-2

Hauk O, Johnsrude I, Pulvermüller F (2004) Somatotopic representation of action words in human motor and premotor cortex. Neuron 41(2):301–307

Ibáñez A, Cardona JF, Dos Santos YV, Blenkmann A, Aravena P, Roca M, Hurtado E, Nerguizian M, Amoruso L, Gómez-Arévalo G, Chade A (2013) Motor-language coupling: direct evidence from early Parkinson's disease and intracranial cortical recordings. Cortex 49 (4):968–984. https://doi.org/10.1016/j.cortex.2012.02.014

Innocenti A, De Stefani E, Sestito M, Gentilucci M (2014) Understanding of action-related and abstract verbs in comparison: a behavioral and TMS study. Cogn Process 15(1):85–92. https://doi.org/10.1007/s10339-013-0583-z

Jacob P, Jeannerod M (2003) Ways of seeing: the scope and limits of visual cognition. Oxford University Press

Jeannerod M (2006) Motor cognition: what actions tell the self. Oxford University Press, Oxford

Kemmerer D, Castillo JG, Talavage T, Patterson S, Wiley C (2008) Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. Brain Lang 107(1):16–43. https://doi.org/10.1016/j.bandl.2007.09.003

Kemmerer D, Rudrauf D, Manzel K, Tranel D (2012) Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. Cortex 48(7):826–848. https://doi.org/10.1016/j.cortex.2010.11.001

Klepp A, Niccolai V, Sieksmeyer J, Arnzen S, Indefrey P, Schnitzler A, Biermann-Ruben K (2017) Body-part specific interactions of action verb processing with motor behavior. Behav Brain Res 328:149–158

Leshinskaya A, Caramazza A (2014) Nonmotor aspects of action concepts. J Cogn Neurosci 26 (12):2863–2879. https://doi.org/10.1162/jocn_a_00679

Levy N (2015) Embodied savoir-faire: knowledge-how requires motor representations. Synthese. https://doi.org/10.1007/s11229-015-0956-1

Lindemann O, Stenneken P, van Schie HT, Bekkering H (2006) Semantic activation in action planning. J Exp Psychol Hum Percept Perform 32(3):633–643. https://doi.org/10.1037/0096-1523.32.3.633

Locatelli M, Gatti R, Tettamanti M (2012) Training of manual actions improves language understanding of semantically related action sentences. Front Psychol 3. https://doi.org/10.3389/fpsyg.2012.00547

Mahon BZ, Caramazza A (2008) A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J Physiol-Paris 102(1–3):59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004

Mele A (1992) Springs of action. Oxford University Press, Oxford

Nanay B (2013) Between perception and action. OUP Oxford, Oxford

Mylopoulos M, Pacherie E (2016) Intentions and motor representations: the interface challenge. Rev Philos Psychol. https://doi.org/10.1007/s13164-016-0311-6

Nazir TA, Boulenger V, Roy A, Silber B, Jeannerod M, Paulignan Y (2008) Language-induced motor perturbations during the execution of a reaching movement. Q J Exp Psychol 61 (6):933–943. https://doi.org/10.1080/17470210701625667

Pacherie E (2000) The content of intentions. Mind Lang 15:400–432

Pacherie E (2008) The phenomenology of action: a conceptual framework. Cognition 107 (1):179–217. https://doi.org/10.1016/j.cognition.2007.09.003

Pacherie E (2011) Non-conceptual representations for action and the limits of intentional control. Soc Psychol 42(1):67–73

Pulvermüller F (2013) Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. Brain Lang 127(1):86–103. https://doi.org/10.1016/j.bandl.2013.05.015

Rizzolatti G, Camarda R, Fogassi L et al (1988) Functional organization of inferior area 6 in the macaque monkey. Exp Brain Res 71:491

Rueschemeyer S-A, Lindemann O, van Rooij D, van Dam W, Bekkering H (2010) Effects of intentional motor actions on embodied language processing. Exp Psychol 57(4):260–266. https://doi.org/10.1027/1618-3169/a000031

Ryle G (1949) The concept of mind. University of Chicago Press

Searle JR (1983) Intentionality: an essay in the philosophy of mind. Cambridge University Press, New York

Shepherd J (2017) Skilled action and the double life of intention. Philos Phenomenol Res 1–20. https://doi.org/10.1111/phpr.12433

Sinigaglia C, Butterfill SA (2015) On a puzzle about relations between thought, experience and the motoric. Synthese 192(6):1923–1936. https://doi.org/10.1007/s11229-015-0672-x

Stanley J, Williamson T (2001) Knowing how. J Philos 98(8):411–444. https://doi.org/10.2307/2678403

Stanley J (2011) Know how. OUP Oxford, Oxford

Tettamanti M, Buccino G, Saccuman MC, Gallese V, Danna M, Scifo P, Fazio F, Rizzolatti G, Cappa SF, Perani DJ (2005) Listening to action-related sentences activates fronto-parietal motor circuits, J Cogn Neurosci 17(2):273–281

Tomasino B, Maieron M, Guatto E, Fabbro F, Rumiati RI (2013) How are the motor system activity and functional connectivity between the cognitive and sensorimotor systems modulated by athletic expertise? Brain Res 1540:21–41. https://doi.org/10.1016/j.brainres.2013.09.048

van Dam WO, van Dongen EV, Bekkering H, Rueschemeyer S-A (2012) Context-dependent changes in functional connectivity of auditory cortices during the perception of object words. J Cogn Neurosci 24(10):2108–2119. https://doi.org/10.1162/jocn_a_00264

van Elk M, van Schie HT, Bekkering H (2008) Conceptual knowledge for understanding other's actions is organized primarily around action goals. Exp Brain Res 189(1):99–107. https://doi.org/10.1007/s00221-008-1408-7

van Elk M, van Schie HT, Zwaan RA, Bekkering H (2010) The functional role of motor activation in language processing: motor cortical oscillations support lexical- semantic retrieval. NeuroImage 50(2):665–677. https://doi.org/10.1016/j.neuroimage.2009.12.123

Willems RM, Hagoort P, Casasanto D (2010) Body-specific representations of action verbs: neural evidence from right- and left-handers. Psychol Sci 21(1):67–74. https://doi.org/10.1177/0956797609354072

Wu H, Mai X, Tang H, Ge Y, Luo Y-J, Liu C (2013) Dissociable somatotopic representations of chinese action verbs in the motor and premotor cortex. Sci Rep 3. https://doi.org/10.1038/srep02049

Zipoli Caiani S, Ferretti G (2017) Semantic and pragmatic integration in vision for action. Conscious Cogn 48:40–54. https://doi.org/10.1016/j.concog.2016.10.009



A Computational-Hermeneutic Approach for Conceptual Explicitation

David Fuenmayor[1(✉)] and Christoph Benzmüller[1,2]

[1] Freie Universität Berlin, Berlin, Germany
david.fuenmayor@fu-berlin.de
[2] University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract. We present a computer-supported approach for the logical analysis and conceptual explicitation of argumentative discourse. Computational hermeneutics harnesses recent progresses in automated reasoning for higher-order logics and aims at formalizing natural-language argumentative discourse using flexible combinations of expressive non-classical logics. In doing so, it allows us to render explicit the tacit conceptualizations implicit in argumentative discursive practices. Our approach operates on networks of structured arguments and is iterative and two-layered. At one layer we search for logically correct formalizations for each of the individual arguments. At the next layer we select among those correct formalizations the ones which honor the argument's dialectic role, i.e. attacking or supporting other arguments as intended. We operate at these two layers in parallel and continuously rate sentences' formalizations by using, primarily, inferential adequacy criteria. An interpretive, logical theory will thus gradually evolve. This theory is composed of meaning postulates serving as explications for concepts playing a role in the analyzed arguments. Such a recursive, iterative approach to interpretation does justice to the inherent circularity of understanding: the whole is understood compositionally on the basis of its parts, while each part is understood only in the context of the whole (hermeneutic circle). We summarily discuss previous work on exemplary applications of human-in-the-loop computational hermeneutics in metaphysical discourse. We also discuss some of the main challenges involved in fully-automating our approach. By sketching some design ideas and reviewing relevant technologies, we argue for the technological feasibility of a highly-automated computational hermeneutics.

Keywords: Computational philosophy · Higher-order logic · Theorem proving · Logical analysis · Hermeneutics · Explication

Benzmüller received funding for this research from VolkswagenStiftung under grant CRAP 93678 (Consistent Rational Argumentation in Politics).

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 441–469, 2019.
https://doi.org/10.1007/978-3-030-32722-4_25

> "... that the same way that the whole is, of course, understood in reference to the individual, so too, the individual can only be understood in reference to the whole."
>
> Friedrich Schleiermacher (1829)

1 Introduction

Motivated by, and reflecting upon, previous work on the computer-supported assessment of challenging arguments in metaphysics (e.g. [12,13,33]), we have engaged in the development of a systematic approach towards the logical analysis of natural-language arguments amenable to (partial) automation with modern theorem proving technology. In previous work [35,36] we have presented some case studies illustrating a computer-supported approach, termed *computational hermeneutics*, which has the virtue of addressing argument formalization and assessment in a holistic way: The adequacy of candidate formalizations for sentences in an argument is assessed by computing the logical validity of the argument as a whole (which depends itself on the way we have so far formalized all of its constituent sentences). Computational hermeneutics has been inspired by ideas in the philosophy of language such as *semantic holism* and Donald Davidson's *radical interpretation* [27]. It is grounded on recent progresses in the area of automated reasoning for higher-order logics and integrates techniques from argumentation theory [36].

Drawing on the observation that "every formalization is an interpretation", our approach has initially aimed at fostering human understanding of argumentative (particularly theological) discourse [7,35]. In the following sections, aside from presenting a more detailed theoretical account of our approach, we want to explore possibilities for further automation, thus gradually removing the human from the loop. In Sect. 2 we introduce the notions of *ontology* and *conceptualization* as defined in the fields of knowledge engineering and artificial intelligence. We use this terminological framework to discuss what it means for us to claim that a logical theory can serve to *explicitate a conceptualization*. In Sect. 3 we discuss what makes our proposed approach hermeneutical. We briefly present a modified account of Donald Davidson's theory of radical interpretation and relate it to the logical analysis of natural-language arguments, in particular, to the problem of assessing the adequacy of sentences' formalizations in the context of an argument (network). We show how an interpretive process grounded on the logical analysis of natural-language argumentative discourse will exhibit a (virtuous) circularity, and thus needs to be approached in a mixed recursive–iterative way.

We also illustrate the fundamental role of computers in supporting this process. In particular, by automating the computation of inferential adequacy criteria of formalization, modern theorem proving technology for higher-order logics can provide the effective and reliable feedback needed to make this approach a viable alternative. In Sect. 4 we briefly present an exemplary case study involving the analysis of a metaphysical argument and discuss some technologies and

implementation approaches addressing some of the main challenges concerning a highly-automated computational hermeneutics. In particular, we discuss a key technique termed *semantical embeddings* [4,10], which harnesses the expressive power of classical higher-order logic to enable a truly logical-pluralist approach towards representing and reasoning with complex theories by reusing state-of-the-art (higher-order) automated reasoning infrastructure.

2 Explicitating Conceptualizations

Computational hermeneutics helps us make our *tacit* conceptualizations *explicit*. In this section we aim at building the necessary background to make best sense of this statement. We start with a cursory presentation of the notion of *an* ontology as used in the fields of artificial intelligence and knowledge engineering. We introduce some of the definitions of (an) ontology drawing on the concept of conceptualization (this latter notion being the one we are mostly interested in). We then have a deeper look into the notion of conceptualizations and how they can be explicitly represented by means of a logical theory.

2.1 Ontologies and Meaning Postulates

Ontology is a quite overloaded term, not only in its original philosophical scope, but also in computer science. Trying to improve this situation, researchers have made the apt distinction between "Ontology" and "an ontology" [42]. The term "Ontology" refers to the philosophical field of study, which will not be considered in this paper. Regarding the latter term ("an ontology") several authoritative definitions have been put forward during the nineties. We recall some of them: Tom Gruber originally defines an ontology as "an explicit specification of a conceptualization" [41]. This definition is further elaborated by Studer et al. [54] as "a formal, explicit specification of a shared conceptualization", thus emphasizing the dimensions of intersubjective conspicuousness and representability by means of a formal language. Nicola Guarino, a pioneer advocate of ontologies in computer science, has depicted an ontology as "a logical theory which gives an explicit, partial account of a conceptualization" [42], thus emphasizing an aspect of insufficiency: an ontology can only give a *partial* account of a conceptualization.

These definitions duly highlight the aspect of explicitness: By means of the articulation in a formal language, an ontology can become common and conspicuous enough to fulfill its normative role—as public standard—for the kind of systems developed in areas like knowledge engineering and, more recently, the Semantic Web. More technically, we see that an ontology (at least in computer science) can be aptly considered as a kind of logical theory, i.e. as a set of sentences or formulas. The question thus arises: Which particular formal properties should the sentences comprising a logical theory have in order to count as an ontology? To be sure, the logical theory may feature some particular annotation or indexing schema distinguishing 'ontological' sentences from 'non-ontological'

ones. But other than this there is no clear line outlining the 'ontological' sentences in a logical theory.[1]

This issue is reminiscent of the controversy around the old analytic–synthetic distinction in philosophy. We do not want to address this complex debate here. However, we want to draw attention to a related duality introduced by Carnap [24] in his notion of *meaning postulates* (and its counterpart: *empirical postulates*). Carnap's position is that, in order to rigorously specify a language—or a logical theory, for our purposes—one is confronted with a *decision* concerning which sentences are to be taken as analytic, i.e. which ones we should attach the label "meaning postulate". For Carnap, meaning postulates are axioms of a definitional nature, which tell us how the meanings of terms are interrelated.[2] They are warranted by an *intent to use* those terms in a certain way (e.g. to draw some inferences that we accept as valid) rather than by any appeal to facts or observations. In this sense, meaning postulates are to be distinguished from factual assertions, also termed "empirical postulates". A kind of analytic–synthetic distinction is thus made, but this time by an appeal to pragmatical considerations. Whether a sentence such as "No mammals live in water" is analytic, or not, depends upon a decision about how to *use* the corresponding concepts in a certain area of discourse; e.g. we may want to favor some inferences and disfavor others (e.g. excluding whales from being considered as mammals). We thus start to see how a listing of meaning postulates can give us some insight into the conceptualization underlying a discourse. Following Carnap, we distinguish in our approach between meaning and empirical postulates. The former are sets of axioms telling us how the meanings of terms are interrelated, and thus constitute our interpretive logical theory. The latter correspond to formalized premises and conclusions of arguments which are to become validated in the context of our theory.

As discussed above, it is rather the intended purpose which lets us consider a logical theory (or a part thereof) as an ontology. We will thus conflate talk of formal ontologies with talk of logical theories for the sake of illustrating our computational-hermeneutic approach.

[1] This is similar to the distinction between *TBox* and *ABox* in knowledge bases. Some may claim that 'ontological' sentences (TBox) tend to be more permanent and mostly concern types, classes and other universals; while other, 'non-ontological' sentences (ABox) mostly concern their instances. The former may be treated as being always true and the latter as subject to on-line revision. However, what counts as a class, what as an instance and what is subject to revision is heavily dependent on the use we intend to give to the theory (knowledge base).

[2] Recalling Carnap's related notion of *explication* [23], we can think of a set of meaning postulates as providing a precise characterization for some new, exact concept (*explicatum*) aimed at "replacing" an inexact, pre-theoretical notion (*explicandum*), for the purpose of advancing a theory. Thus, in computational hermeneutics, the non-logical terms of our interpretive theory characterize concepts playing the role of *explicata* aimed at explicitly representing fuzzy, unarticulated *explicanda* from a tacit conceptualization.

2.2 Conceptualizations

As seen from the above, definitions of (an) ontology depict a conceptualization, by contrast, as something tacit and unarticulated; as something being *hinted at* by the use of natural language, instead of being *determined by* the semantics of a formal one. In the following, we want to evoke this connotation every time we use the term conceptualization *simpliciter*. Our notion of conceptualization (*simpliciter*) thus refers to that kind of implicit, unarticulated and to some extent undetermined knowledge being framed by human socio-linguistic practices, and which is to become (partially) represented—and thereby *made explicit*—by means of a logical theory. In particular, we envisage a certain special—and arguably primordial—kind of socio-linguistic practice: argumentative discourse.

We have a formal definition for the notion of conceptualization starting with Genesereth and Nilsson [39], who state that "the formalization of knowledge in declarative form begins with a *conceptualization*. This includes the objects presumed or hypothesized to exist in the world and their interrelationships." In their account, objects "can be anything about which we want to say something." Genesereth and Nilsson then proceed to formally define a conceptualization as an *extensional* relational structure: a tuple $\langle D, R \rangle$ where D is a set called the universe of discourse and R is a collection of relations (as sets of tuples) on D. This account of a conceptualization has been called into question because of its restricted extensional nature: A conceptualization, so conceived, concerns only how things in the universe of discourse *actually are* and not how they *could acceptably be* interrelated. The structure proposed by Genesereth and Nilsson as a formal characterization of the notion of conceptualization seems to be more appropriate for states of affairs or world-states; a conceptualization having a much more complex (social or mental) nature.

Other, more expressive definitions for the notion of conceptualization have followed. Gruber [41] states that a conceptualization is "an abstract, simplified view of the world that we wish to represent for some purpose"; and Uschold [57] sees a conceptualization as a "world-view", since it "corresponds to a way of thinking about some domain". According to the account provided by Guarino and his colleagues [42,43], a conceptualization "encodes the implicit rules constraining the structure of a piece of reality". Note that this characterization, in contrast to the one provided by Genesereth and Nilsson, does not talk about how a "piece of reality" actually is, but instead about how it can possibly or acceptably be, according to some constraints set by *implicit* rules.[3] Guarino has also provided a formal account of conceptualizations, which we reproduce below (taken from [43]). In the following, we distinguish the informal notion of conceptualization *simpliciter* from the formal, exact concept introduced below. Note that we will always refer to the latter in a qualified form as a *formal* conceptualization.

[3] As we see it, those rules are (their tacitness notwithstanding) of a logical nature: they concern which arguments or inferences are endorsed by (a community of) speakers.

Definition 1 (Intensional relational structure, or formal conceptualization). *An intensional relational structure (or a conceptualization according to Guarino) is a triple* $C = \langle D, W, \mathbb{R} \rangle$ *with*

- D *a universe of discourse, i.e. an arbitrary set of individual entities.*
- W *a set of possible worlds (or world-states). Each world is a maximal observable state of affairs, i.e. a unique assignment of values to all the observable variables that characterize the system.*
- $\mathbb{R}$ *a set of intensional (aka. conceptual) relations. An intensional relation is a function mapping possible worlds to extensional relations on* D*.*

Taking inspiration from Carnap [23] and Montague [28], Guarino takes conceptual (intensional) relations to be functions from worlds to extensional relations (sets of tuples). This account aims at doing justice to the intuition that conceptualizations—being arguably about concepts—should not change when some particular objects in the world become otherwise related. However, *formal conceptualizations*, while having the virtue of being exactly defined and unambiguous, are just fictional objects useful to clarify the notion of an ontology.

2.3 Representing a Conceptualization by Means of a Theory

For Guarino and his colleagues, the role of formal conceptualizations in the definition of an ontology is that of a touchstone. For them an ontology is—let us recall their definition—"a logical theory which gives an explicit, partial account of a conceptualization" [42]. Thus, an ontology—or more specifically: its models[4]—can only *partially* represent a conceptualization (formal or *simpliciter*). In order to evaluate how good a given ontology represents a conceptualization, Guarino has introduced a formal notion of *ontological commitment* (similar in spirit to the well-known eponymous notion in philosophy). According to Guarino, by using a certain language a speaker commits—even unknowingly—to a certain conceptualization. Such a commitment arises from the fact that, by employing a certain linguistic expression (e.g. a name or an adjective), a language user *intends* to refer to objects which are part of a conceptualization, i.e. some individual or (conceptual) relation. Those objects may be tacit in the sense that they become presupposed in our inferential practices.[5]

[4] When we talk of models of a logical theory or ontology, we always refer to models in a model-theoretical sense, i.e. interpretations: assignments of values (as denoted entities) to non-logical terms.

[5] For instance, the existence of human races may need to be posited for some eugenicist arguments to succeed; or the presupposition of highly-localized specific brain functions may be needed for a phrenology-related argument to get through. When we intuitively accept the conclusions of arguments we may thereby also commit to the existence of their posits (as made explicit in the logical forms of *adequate* formalizations). Conversely, such arguments can be attacked by calling into question the mere existence of what they posit.

Guarino's account leaves some room for partiality (incompleteness) in our representations. By interpreting natural language, suitably represented as a logical theory, an interpreter can indeed end up referring to entities other than those originally intended or conceiving them in undesired ways. In such case we would say that some of the models (interpretations) of the theory are not among its intended models (see Fig. 1). In Guarino's view, what is intended becomes prescribed by the target conceptualization. For instance—to put it in informal terms—the conceptualization I commit to could prescribe that all fishes live in water. However, in one of the possible interpretations (models) of some theory of us (e.g. about pets), which includes the sentence "Nemo lives in the bucket", the term "Nemo" may refer to an object falling under the "fish" predicate and the expression "the bucket" may refer to some kind of waterless cage. If, in spite of this rather clumsy interpretation, the theory is still consistent, we would say that the theory somehow 'underperforms': it does not place enough constraints as to preclude undesired interpretations, i.e. has models which are not intended. Underperforming theories (ontologies) in this sense are more the rule than the exception. Now let us put the previous discussion in formal terms (taken from Guarino et al. [43]):

Definition 2 (Extensional relational structure). *An extensional relational structure, (or a conceptualization according to Genesereth and Nilsson [39]), is a tuple* $\langle D, R\rangle$ *where* D *is a set called the universe of discourse and* R *is a set of (extensional) relations on* D.

Definition 3 (Extensional first-order structure, or model). *Let* L *be a first-order logical language with vocabulary* V *and* $S = \langle D, R\rangle$ *an extensional relational structure. An extensional first-order structure (also called model for* L*) is a tuple* $M = \langle S, I\rangle$*, where* I *(called extensional interpretation function) is a total function* $I : V \rightarrow D \cup R$ *that maps each vocabulary symbol of* V *to either an element of* D *or an extensional relation belonging to the set* R.

Definition 4 (Intensional first-order structure, or ontological commitment). *Let* L *be a first-order logical language with vocabulary* V *and* $C = \langle D, W, \mathbb{R}\rangle$ *an intensional relational structure (i.e. a conceptualization). An intensional first-order structure (also called ontological commitment) for* L *is a tuple* $K = \langle C, \mathbb{I}\rangle$*, where* $\mathbb{I}$ *(called intensional interpretation function) is a total function* $\mathbb{I} : V \rightarrow D \cup \mathbb{R}$ *that maps each vocabulary symbol of* V *to either an element of* D *or an intensional relation belonging to the set* $\mathbb{R}$.

Definition 5 (Intended models of a theory w.r.t. a formal conceptualization). *Let* $C = \langle D, W, \mathbb{R}\rangle$ *be a conceptualization,* L *a first-order logical language with vocabulary* V *and ontological commitment* $K = \langle C, \mathbb{I}\rangle$*. A model* $M = \langle S, I\rangle$*, with* $S = \langle D, R\rangle$*, is called an intended model of* L *according to* K *iff*

1. *For all constant symbols* $c \in V$ *we have* $I(c) = \mathbb{I}(c)$
2. *There exists a world* $w \in W$ *such that, for each predicate symbol* $v \in V$ *there exists an intensional relation* $\rho \in \mathbb{R}$ *such that* $\mathbb{I}(v) = \rho$ *and* $I(v) = \rho(w)$

The set $I_K(L)$ of all models of L that are compatible with K is called the set of intended models of L according to K.

Definition 6 (An ontology). *Let C be a conceptualization, and L a logical language with vocabulary V and ontological commitment K. An ontology O_K for C with vocabulary V and ontological commitment K is a logical theory consisting of a set of formulas of L, designed so that the set of its models approximates the set $I_K(L)$ of intended models of L according to K* (see Fig. 1).

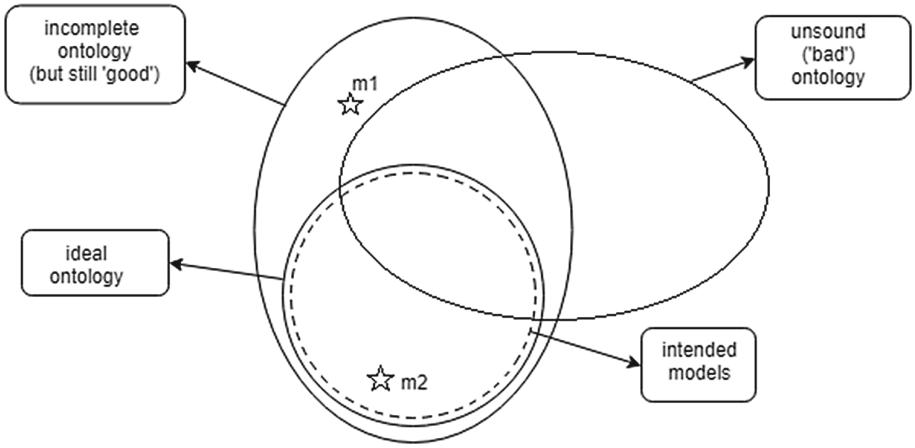


Fig. 1. Model "m1" is a non-intended model of the 'good' yet incomplete ontology; whereas "m2" is an intended model left out by the unsound, 'bad' ontology.

Given the definitions above we might conclude, following Guarino, that an ideal ontology is one whose models exactly coincide (modulo isomorphism) with the intended ones. Variations to this ideal would entail (with respect to a formal conceptualization) either (i) incompleteness: the ontology has models that are non-intended, and thus truth in the conceptualization does not entail validity in the ontology (logical theory); or (ii) unsoundness: the ontology rules out intended models. This latter situation is the most critical, since the ontology *qua logical theory* would license inferences that are not valid in the target conceptualization.[6] (See Fig. 1 for a diagrammatic illustration.)

Back to the idea of computational hermeneutics, we concede that there are no means of mechanically verifying coherence with a conceptualization *simpliciter*, the latter being something tacit, fuzzy and unarticulated. However, we do have the means of deducing and checking consequences drawn from a logical theory, and this is indeed the reason why computational hermeneutics relies on

[6] Guarino further considers factors like language expressivity (richness of logical and non-logical vocabulary) and scope of the domain of discourse in having a bearing on the degree to which an ontology specifies a conceptualization.

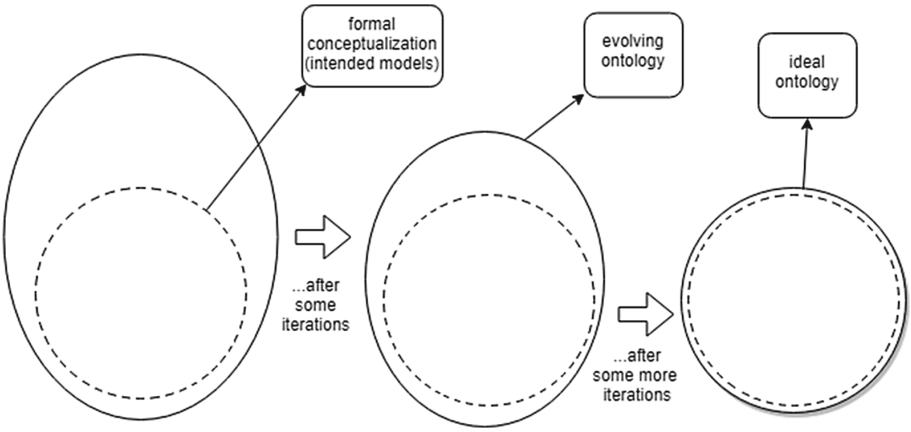


Fig. 2. An idealized, straightforward interpretive approach.

the use of automated theorem proving. Regarding formal conceptualizations, we may have the means of mechanically verifying coherence with them (since they are well-defined mathematical structures), but this would not bring us much further. As mentioned before, a formal conceptualization is a fictional mathematical object, which might at best serve as a model (in the sense of being an approximated representation) for some real—though unarticulated—conceptualization. The notion of formal conceptualization has been brought forth in order to enable the previous theoretical analysis and foster understanding. Hopefully, we can now make better sense of the claim that computational hermeneutics works by iteratively evolving a logical theory towards adequately approximating a conceptualization, thus making the latter explicit.

2.4 An Idealized Interpretive Approach

As illustrated above, in artificial intelligence and knowledge engineering, ontologies correspond to logical theories and introduce a vocabulary (logical and non-logical) together with a set of formalized sentences (formulas). Some of these formulas: axioms, directly aim at ruling out unintended interpretations of the vocabulary so that as many models as possible correspond to those worlds (i.e. maximal states of affairs) compatible with our target (formal) conceptualization. Theorems also do this, but indirectly, by constraining axioms (since the former have to follow from the latter). Hence, an ontology *qua logical theory* delimits what is possible 'from the outside', i.e. it tells us which interpretations of the given symbols (vocabulary) are not acceptable with respect to the way we intuitively understand them. Every time we add a new sentence to the theory we are rejecting those symbols' interpretations (i.e. models) which render the (now augmented) theory non-valid. Conversely, when removing a sentence some interpretations (models) become acceptable again. We can thus think of a mechanism to iteratively evolve a logical theory by adding and removing axioms and

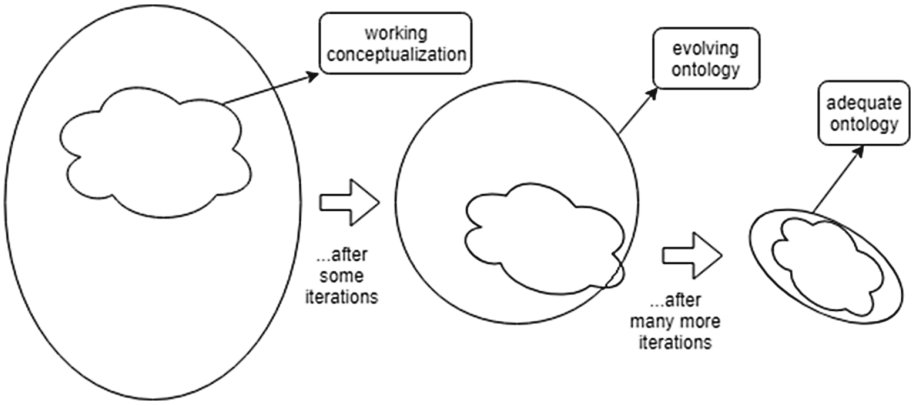


Fig. 3. A more realistic view, where a conceptualization is a moving target.

theorems, while getting appropriate feedback about the adequacy of our theory changes. Such a mechanism would converge towards an optimal solution insofar as we can get some kind of quantitative feedback regarding how well the models of our theory are approximating the intended ones (i.e. the target formal conceptualization). As discussed more fully below, this is a quite simplistic assumption. However, it is still worth depicting for analysis purposes as we did for (fictional) formal conceptualizations. This is shown in Fig. 2.

As mentioned before, *formal* conceptualizations do not exist in real life. We do not start our interpretive endeavors with a listing of all individual entities in the domain of discourse together with the relations that may acceptably hold among them. We have therefore no way to actually determine which are our intended models. In the situation depicted in Fig. 2 we were chasing a ghost. Moreover—to make things more interesting—our conceptualization indeed changes as we move through our interpretive process: We learn more about the concepts involved and about the consequences of our beliefs. We can even call some of them into question and change our minds. Figure 3 shows a more realistic view of the problem, where we are chasing a moving target.

However, not all hope is lost. In the context of argumentative discourse, we still believe in the possibility of devising a feedback mechanism which can account for the adequacy of our interpretive logical theories. Instead of contrasting sets of models with formal conceptualizations, we will be putting sets of formalized sentences and (extracts from) natural-language arguments side-by-side and then comparing them among different dimensions. Thus, we will be reversing the account given by Guarino and his colleagues, by arguing that conceptualizations first originate (i.e. come into existence in an explicit and articulated way) in the process of developing an interpretive theory. Conceptualizations correspond with 'objective reality' to a greater or lesser extent as they license inferential moves which are endorsed by a linguistic community in the context of some

argumentative discourse (seen as a network of arguments). The diagram shown in Fig. 4 gives a general idea of our interpretive approach.

3 A Hermeneutic Approach

Argumentative discourse and conceptualizations are two sides of a coin. In the previous section we briefly mentioned the idea of putting logical theories and natural-language arguments side-by-side and then comparing them among different dimensions. One of those dimensions—indeed, the main one—is *truth*, or more appropriately for arguments: *logical correctness*.[7] This idea has been inspired by Donald Davidson philosophy of language, in particular, his notion of *radical interpretation*. In this section we aim at showing how the problem of formalizing argumentative discourse, understood as networks of arguments

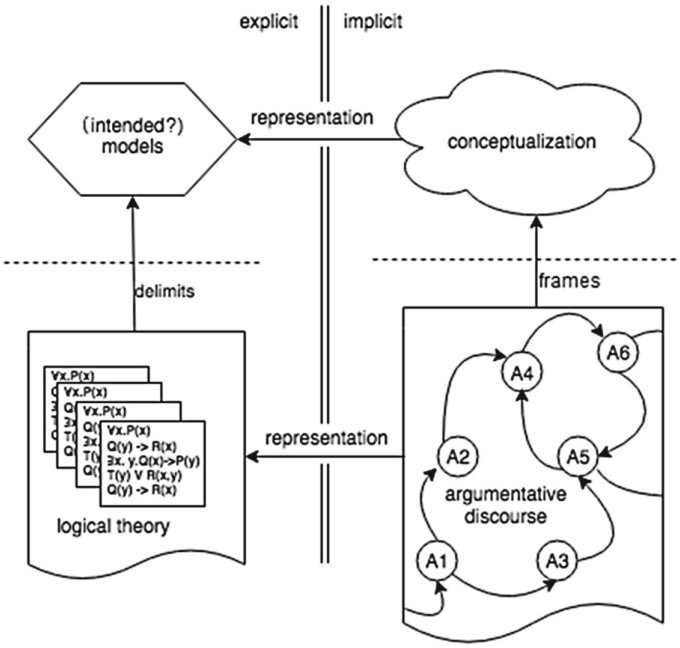


Fig. 4. How (models of) logical theories represent conceptualizations.

[7] Logical correctness encompasses, among others, the more traditional concept of logical validity. Our working notion of logical correctness also encompasses axioms/premises consistency and lack of circularity (no *petitio principii*) as well as avoiding idle premises. Other accounts may consider different criteria. We have restricted ourselves to the ones that can be efficiently computed with today's automated reasoning technology. See [50] for an interesting discussion of logical (in)correctness.

mutually attacking and supporting each other, relates to the problem of adequately representing a conceptualization. We share with the supporters of logical expressivism (see e.g. [18,48]) the view that logical theories are means to explicitly articulate the rules implicit in discursive practices.

3.1 Hermeneutic Circle

Previously, we hinted at the possibility of a feedback mechanism that can account for the adequacy of our interpretive logical theories. We commented on the impossibility of directly contrasting the models of our theories with the target conceptualization (determining the theory's intended models), which would have given us a straightforward search path for the 'best' theory (see Fig. 2). The main difficulty is that, aside from their complex tacit nature, conceptualizations are also a moving target (as shown in Fig. 3). Our conceptualizations are work in progress and continually change as we take part in socio-linguistic practices, especially argumentation. We may thus call them *working* conceptualizations, since they are invariably subject to revision. This can be seen both at the individual and societal level. For instance, when we as understanding agents become more aware of (or even learn something new about) our concepts and beliefs. As a community of speakers, through interpretive reconstruction of public discourse, we can become more self-conscious of our tacit conceptualizations as framed by prevalent discursive practices, thus enabling their critical assessment and eventual revision.[8]

Adhering to the slogan: "Every formalization is an interpretation", the logical analysis of arguments becomes essentially an interpretive endeavor. A characteristic of this endeavor is its recursive, holistic nature. The logical forms—as well as the meanings—of the individual statements comprising an argument (network) are holistically interconnected. At any point in our interpretive process, the adequacy of the formalization of some statement (sentence) will hinge on the way other related sentences have been formalized so far, and thus depend on the current state of our interpretive theory, i.e. on our *working* conceptualization. Donald Davidson aptly illustrates this situation in the following passage [26, p. 140].

> "... much of the interest in logical form comes from an interest in logical geography: to give the logical form of a sentence is to give its logical location in the totality of sentences, to describe it in a way that explicitly determines what sentences it entails and what sentences it is entailed by. The location must be given relative to a specific deductive theory; so logical form itself is relative to a theory."

[8] Sharing a similar background and motivation, computational hermeneutics might support a technological implementation of contemporary approaches for the *revisionary* philosophical analysis of public discourse like ameliorative analysis (in particular, as presented in [47]), conceptual ethics [22], and conceptual engineering (e.g. as discussed in [20]).

We thus conceive this hermeneutic interpretive endeavor as a holistic, iterative enterprise. We start with tentative, simple candidate formalizations of an argument's statements and iteratively use them as stepping stones on the way to the (improved) formalization of others. The adequacy of any sentence's formalization becomes tested by evaluating the logical correctness of the formalized arguments (from which it is a component). Importantly, evaluating logical correctness is a holistic operation that involves all of an argument's constituent sentences. That is, the result of any formalization's adequacy test becomes dependent not only upon the choice of the particular formula that is put to the test, but also upon previous formalization choices for its companion sentences in the arguments composing a discourse (i.e. an argument network). Moreover, the formalization of any individual sentence depends (by compositionality) on the way its constituent terms have been explicated (using meaning postulates). Thus, our approach does justice to the inherent circularity of interpretation, where the whole is understood compositionally on the basis of its parts, while each part can only be understood in the context of the whole. In the philosophical literature (particularly in [37]) this recursive nature of interpretation has been termed: the *hermeneutic circle*.

3.2 Radical Interpretation

By putting ourselves in the shoes of an interpreter aiming at 'translating' some natural-language argument into a formal representation, we have had recourse to Donald Davidson's philosophical theory of *radical interpretation* [27]. Davidson builds upon Quine's account of *radical translation* [51], which is an account of how it is *possible* for an interpreter to understand someone's words and actions without relying on any prior understanding of them.[9] Since the radical interpreter cannot count on a shared language for communication nor the help of a translator, she cannot directly ask for any kind of explanations. She is thus obliged to conceive interpretation hypotheses (theories) and put them to the test by gathering 'external-world' evidence regarding their validity. She does this by observing the speaker's use of language in context and also by engaging in some basic dialectical exchange with him/her (e.g. by making utterances while pointing to objects and asking yes–no kind of questions).

Davidson's account of radical interpretation builds upon the idea of taking the concept of *truth* as basic and extracting from it an account of interpretation satisfying two general requirements: (i) it must reveal the compositional structure of language, and (ii) it can be assessed using evidence available to the interpreter. The first requirement (i) is addressed by noting that a theory of truth in Tarski's style [56] (modified to apply to natural language) can be used as

[9] In this sense, Davidson has emphatically made clear that he does not aim at showing how humans *actually* interpret let alone acquire natural language. This being rather a subject of empirical research (e.g. in cognitive science and linguistics) [25]. However, Davidson's philosophical point becomes particularly interesting in artificial intelligence, as regards the design of artificial language-capable machines.

a theory of interpretation. This implies that, for every sentence s of some object language L, a sentence of the form: «"s" is true in L iff p» (aka. T-schema) can be derived, where p acts as a translation of s into a sufficiently expressive language used for interpretation (note that in the T-schema the sentence p is being *used*, while s is only being *mentioned*). Thus, by virtue of the recursive nature of Tarski's definition of truth, the *compositional* structure of the object-language sentences becomes revealed. From the point of view of computational hermeneutics, the sentence s is to be interpreted in the context of a given argument (or a network of mutually attacking/supporting arguments). The language L thereby corresponds to the idiolect of the speaker (natural language), and the target language is constituted by formulas of our chosen logic of formalization (some expressive logic XY) plus the meta-logical turnstyle symbol $\vdash_{\mathrm{XY}}$ signifying that an inference (argument or argument step) is valid in logic XY. As an illustration, consider the following instance of the T-schema:

> «"Fishes are necessarily vertebrates" is true [in English, in the context of argument A] iff $A_1, A_2, ..., A_n \ \vdash_{\mathrm{MLS4}}$ "$\forall x.\ \mathit{Fish}(x) \rightarrow \Box\, \mathit{Vertebrate}(x)$"»

where $A_1, A_2, ..., A_n$ correspond to the formalization of the premises of argument A and the turnstyle $\vdash_{\mathrm{MLS4}}$ corresponds to the standard logical consequence relation in the chosen logic of formalization, e.g. a modal logic S4 (MLS4).[10] This toy example aims at illustrating how the interpretation of a statement relates to its logic of formalization and to the inferential role it plays in a single argument. Moreover, the same approach can be extended to argument networks. In such cases, instead of using the notion of logical consequence (represented above as the parameterized logical turnstyle $\vdash_{\mathrm{XY}}$), we can work with the notion of argument *support*. It is indeed possible to parameterize the notions of *support* and *attack* common in argumentation theory with the logic used for argument's formalization [36].

The second general requirement (ii) of Davidson's account of radical interpretation states that the interpreter has access to objective evidence in order to judge the appropriateness of her interpretation hypotheses, i.e., access to the events and objects in the 'external world' that cause statements to be true—or, in our case, arguments to be valid. In our approach, formal logic serves as a common ground for understanding. For instance, computing the logical validity (or a counterexample) of a formalized argument constitutes the kind of objective—or more appropriately: intersubjective—evidence needed to ground the adequacy of our interpretations, under the *charitable* assumption that the speaker (i.e. whoever originated the argument) follows, or at least accepts, similar logical rules as we do. In computational hermeneutics, the computer acts as an (arguably unbiased) arbiter deciding on the correctness of arguments in the context of some encompassing argumentative discourse—which itself tacitly frames the conceptualization that we aim at explicitating.

[10] As will be described in Sect. 4.1, the semantical embeddings approach [4,10] allows us to embed different non-classical logics (modal, deontic, intuitionistic, etc.) in higher-order logic (as meta-language), and to combine them dynamically by adding and removing axioms.

3.3 The Principle of Charity

A central notion in Davidson's account of radical interpretation is the *principle of charity*, which he holds as a condition for the possibility of engaging in any kind of interpretive endeavor. The principle of charity has been summarized by Davidson by stating that "we make maximum sense of the words and thoughts of others when we interpret in a way that optimizes agreement" [27, p. 197]. Hence the principle builds upon the possibility of intersubjective agreement about external facts among speaker and interpreter. The principle of charity can be invoked to make sense of a speaker's ambiguous utterances and, in our case, to presume (and foster) the correctness of an argument. Consequently, in computational hermeneutics we assume from the outset that the argument's conclusions indeed follow from its premises and disregard formalizations that do not do justice to this postulate. Other criteria like avoiding inconsistency and *petitio principii* are also taken into account. At the argument network level, we foster formalizations which honor the intended dialectic role of arguments, i.e. attacking or supporting other arguments as intended.

3.4 Adequacy Criteria of Formalization

We start to see how the problem of formalizing natural-language arguments relates to the problem of explicitating a conceptualization. The interpretive theory that gradually evolves in the holistic, iterative computational-hermeneutic process is composed of those meaning postulates which have become 'settled' during the (many) iterations involving arguments' formalization and assessment: It is a logical theory. Not surprisingly, the quality of our resulting theory will be subordinate to the adequacy of the involved argument formalizations. Since arguments are composed of statements (sentences), the question thus arises: What makes a certain sentence's formalization better than another? This topic has indeed been discussed sporadically in the philosophical literature during the last decades (see e.g. [3,16,19,32,53]). More recently, the work of Peregrin and Svoboda [49,50] has emphasized the role of inferential criteria for assessing the adequacy of argument formalizations. These criteria are quite in tune, because of their inferential nature, with the holistic picture of meaning we present here. Furthermore, they lend themselves to mechanization with today's automated reasoning technologies. We recall these inferential criteria below (from [50]):

(i) The *principle of reliability*: "ϕ counts as an adequate formalization of the sentence S in the logical system L only if the following holds: If an argument form in which ϕ occurs as a premise or as the conclusion is valid in L, then all its perspicuous natural language instances in which S appears as a natural language instance of ϕ are intuitively correct arguments."
(ii) The *principle of ambitiousness*: "ϕ is the more adequate formalization of the sentence S in the logical system L the more natural language arguments in which S occurs as a premise or as the conclusion, which fall into the intended scope of L and which are intuitively perspicuous and correct, are instances

of valid argument forms of S in which ϕ appears as the formalization of S." [50, pp. 70–71].

The first principle (i) can be seen as a kind of 'soundness' criterion: Adequate formalizations must not validate intuitively incorrect arguments. The second principle (ii) has an analogous function to that of 'completeness' criteria: We should always aim at the inferentially most *fruitful* formalization, i.e. the one which renders as logically valid (correct) the highest number of intuitively correct arguments. Peregrin and Svoboda have proposed further adequacy criteria of a more syntactic nature, considering similarity of grammatical structure and simplicity (e.g. number of occurrences of logical symbols). They are not considered in-depth here. Moreover, many of the 'settled' formulas which constitute our interpretive (logical) theory will have no direct counterpart in the original natural-language arguments. They are mostly implicit, unstated premises and often of a definitional nature (meaning postulates). More importantly, Peregrin and Svoboda have framed their approach to logical analysis as a holistic, give-and-take process aiming at reaching a state of *reflective equilibrium*.[11] Our work on computational hermeneutics can be seen, relatively speaking, as sketching a possible technological implementation of Peregrin and Svoboda's (and also Brun's [20]) ideas. We are however careful in invoking the notion of reflective equilibrium for other than illustrative purposes as we strive towards clearly defined and computationally amenable termination criteria for our interpretive process.[12]

Not surprisingly, such a holistic approach involves, even for the simplest cases, a search over a vast combinatoric of candidate formalizations, whose adequacy has to be assessed at least several hundreds of times, particularly if we take the logic of formalization as an additional degree of freedom (as we do). An effective evaluation of the above inferential criteria would thus involve automatically computing the logical validity and consistency of formalized arguments (i.e. proofs). This is the kind of work automated theorem provers and model finders are built for. The recent improvements in automated reasoning technology (in particular for higher-order logics) constitute the main enabling factor for computational hermeneutics, which would otherwise remain unfeasible if attempted manually (e.g. carrying out proofs by hand using natural deduction or tableaux calculi).

[11] The notion of *reflective equilibrium* has been initially proposed by Nelson Goodman [40] as an account for the justification of the principles of (inductive) logic and has been popularized years later in political philosophy and ethics by John Rawls [52] for the justification of moral principles. In Rawls' account, "reflective equilibrium" refers to a state of balance or coherence between a set of general principles and particular judgments (where the latter follow from the former). We arrive at such a state through a deliberative give-and-take process of mutual adjustment between principles and judgments. More recent methodical accounts of reflective equilibrium have been proposed as a justification condition for scientific theories [31] and objectual understanding [2], and as a methodology for conceptual engineering [20].

[12] There are ongoing efforts on our part to frame the problem of finding an adequate interpretive theory as an optimization problem to be approached by appropriate heuristic methods.

3.5 Computational Hermeneutics

As concerns human-in-the-loop computational hermeneutics, the dialectical exchange between interpreter and speaker depicted above takes place between humans and interactive proof assistants (think of a philosopher-logician seeking to translate 'unfamiliar' metaphysical discourse into a 'familiar' logical formalism). At any instant during our iterative, hermeneutic process, we will have an evolving set of meaning postulates and a set of further, disposable formulas. We start each new iteration by extending our working interpretive theory with some additional axioms and theorems corresponding to the candidate formalization of some chosen argument.[13] In other words, our theory has been temporarily extended with additional axioms and (possible) theorems corresponding to the formalizations of an argument's premises and conclusion(s) respectively. This (extended) theory is most likely to *not be* logically correct. It is now our task as (human or machine) interpreters to come up with the missing axioms that would validate it without rendering it inconsistent or question-begging, among others. By doing so, we engage in a dialectical exchange with the computer. The questions we ask concern, among others, the logical correctness (validity, consistency, non-circularity, etc.) of the so extended formal arguments. Harnessing the latest developments in (higher-order) theorem proving, such questions can get answered automatically in milliseconds.

During the process of securing logically correct formalizations for each of the arguments in the network, we continuously select among the collected formulas the ones most *adequate* for our purposes. For this we draw upon the inferential adequacy criteria presented above—eventually resorting to further discretionary qualitative criteria. A subset of the formulas will thus qualify for meaning postulates and are kept for future iterations, i.e. they become 'settled'.[14] Those meaning postulates can be identified with an ontology as described in Sect. 2.1. They introduce semantic constraints aimed at ensuring that any instantiation of the non-logical vocabulary is in line with our working conceptualization (i.e. with our tacit understanding of the real-world counterparts these symbols are intended to denote). This is the way argument analysis helps our interpretive theories evolve

[13] Recall that we think of discourses as networks of mutually supporting/attacking arguments. Each formalized argument can be seen as a collection of axioms and theorems; the latter being intended to logically follow from a combination of the former plus some further axioms of a definitional nature (meaning postulates). This view is in tune with prominent structured approaches to argumentation in artificial intelligence (cf. [14,30]).

[14] As mentioned in Sect. 2.1, there is no definitive criteria for distinguishing meaning postulates from others (cf. ontological vs. non-ontological or TBox vs. ABox sentences). The heuristics for labeling sentences as meaning postulates thus constitute another degree of freedom in our interpretive process, which we address primarily (but not exclusively) by means of inferential adequacy criteria. Moreover, our set of meaning postulates can at some point become inconsistent, thus urging us to mark some of them for controlled removal. In this aspect, our approach resembles reason-maintenance and belief-revision frameworks (cf. [29]).

towards better representing the conceptualization implicit in argumentative discourse.

Removing the human from the loop would, arguably, leave us with an artificial language-capable system. However, to get there we first need to find a replacement for human ingenuity in the most important 'manual' steps currently involved in the human-in-the-loop process: (i) generating a rough and ready initial formalization for arguments (which we may call "bootstrapping"); and (ii) the abductive process of coming up with improved candidate formalizations and implicit premises. Not surprisingly, these are very difficult tasks; however, they do not belong to science-fiction. Current developments in natural language processing and machine learning are important enablers for pursuing this ambitious goal. Another important aspect for both human-in-the-loop and fully-automated computational hermeneutics is the availability of databases of natural-language arguments tagged as either correct or incorrect, as the case may be. After choosing the area of discourse whose conceptualization we are interested in, we proceed to gather relevant arguments from the relevant databases and form an argument network with them.[15] Note that, in contrast to highly quantitative, big-data approaches, computational hermeneutics engages in deep semantic analysis of natural-language, and thus needs far less data to generate interesting results. As previous case studies for human-in-the-loop computational hermeneutics applied to ontological arguments have shown (see e.g. [35,36] and Sect. 4.2 below), working with a handful of sentences already provides interesting characterizations (as meaning postulates) for the metaphysical concepts involved, e.g. (necessary) existence, contingency, abstractness/concreteness, Godlikeness, essence, dependence, among others.

4 Examples and Implementation Approaches

In this section we summarily discuss previous case studies and depict technological solutions for the implementation of computational hermeneutics. We start by highlighting the essential role of *higher-order* automated theorem proving and the related technique of *semantical embeddings*. We then depict a landscape of different tools and technologies which are currently being tested and integrated to pursue an (increasingly) automated computational hermeneutics.

4.1 Semantical Embeddings

Our focus on theorem provers for higher-order logics is motivated by the notion of *logical pluralism*: the view that logics are theories of argument's validity, where

[15] Argument databases and arguments extracted from text sources usually provide information on support and attack relations (see [21,44] and references therein). Another alternative is to dynamically construct the needed arguments by using the working theory plus hypothetical premises and conclusions as building stones. Those arguments would then be presented, in an interactive way, to the user for rejection or endorsement. This mode of operation would correspond to a kind of inverted (we could call it 'Socratic') question-answering system.

different yet *legitimate* logics may disagree about which argument forms are valid. In order to pursue a truly logical-pluralist approach in computational hermeneutics, it becomes essential that the (candidate) logics of formalization also form part of our evolving logical theory. That is, the logic of formalization needs to find its way into the sentences constituting our interpretive theory. More specifically, the (meta-logical) definitions for the logical vocabulary used to encode argument's sentences become a further set of axioms which can also be iteratively varied, added and removed as we go. Thus, computational hermeneutics targets the flexible combination of different kinds of classical and non-classical logics (modal, temporal, deontic, intuitionistic, etc.) through the technique of *semantical embeddings* [4,10].

The semantical embeddings approach[16] harnesses the expressive power of classical higher-order logic (HOL)—also Church's type theory [5]—as a *meta-language* in order to embed the syntax and semantics of another logic as an object language, thereby turning theorem proving systems for HOL into universal reasoning engines [4]. HOL is a logic of functions formulated on top of the simply typed λ-calculus, which also provides a foundation for functional programming. The semantics of HOL is well understood [6]. A variant of this approach, termed *shallow* semantical embeddings (SSE), involves the definition of the logical vocabulary of a target logic in terms of the non-logical vocabulary (lambda expressions) of our expressive meta-language (HOL). This technique has been successfully implemented using the interactive proof assistant Isabelle [46].

SSE has been exploited in previous work for the evaluation of arguments in philosophy, particularly in metaphysics [12,13,33,35,36]. The formal representation of metaphysical arguments poses a challenging task that requires the combination of different kinds of expressive higher-order non-classical logics, currently not supported by state-of-the-art automated reasoning technology. In particular, two of the most well-known higher-order proof assistants: *Isabelle* and *Coq* do not support off-the-shelf reasoning with modal logics. Thus, in order to turn Isabelle into a flexible modal logic reasoner we have adopted the SSE approach. Isabelle's logic (HOL) supports the encoding of sets via their characteristic functions represented as λ-terms. In this sense, HOL comes with a in-built notion of (typed) sets that is exploited in our work for the explicit encoding of the truth-sets that are associated with the formulas of higher-order modal logic.[17]

As illustrated in Fig. 5 (lines 21–23), we can embed a modal logic in HOL by defining the modal $\Box$ and $\Diamond$ operators as meta-logical predicates in HOL and using quantification over sets of objects of a definite type w, representing the type of possible worlds or world-states. Formula $\Box\varphi$, for example, is modeled as an abbreviation (syntactic sugar) for the truth-set $\lambda w.\forall v.w\ R\ v \longrightarrow \varphi v$, where

[16] Note that this approach is not related to the similarly named notion of word embeddings in natural language processing (NLP).

[17] Note that since Isabelle-specific extensions of HOL (except for prefix polymorphism) are not exploited in our work, the technical framework we depict here can easily be transferred to other HOL theorem proving environments.

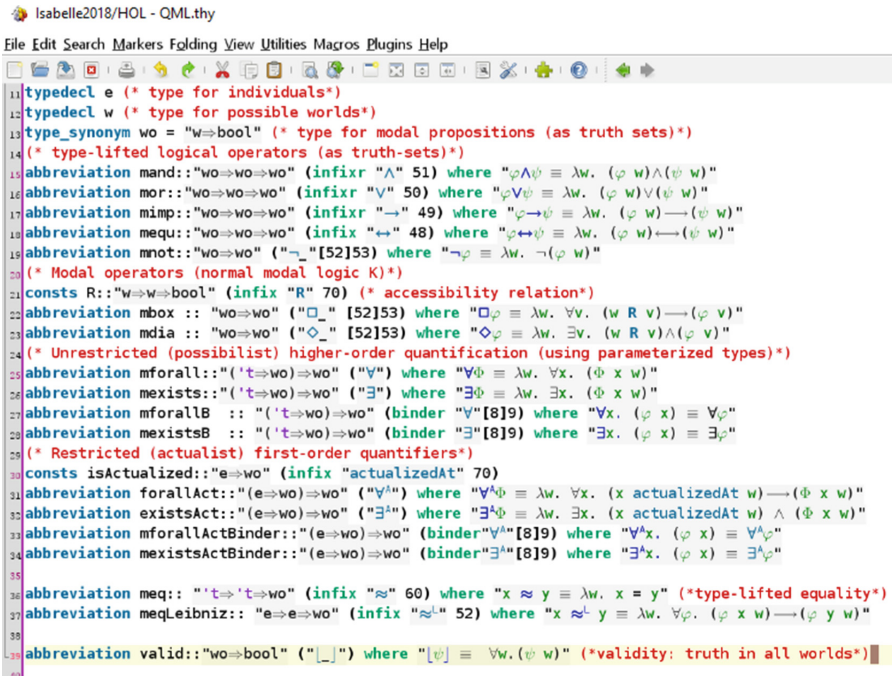


Fig. 5. SSE of a higher-order modal logic in Isabelle/HOL.

R denotes the accessibility relation associated with the modal $\Box$ operator. All presented equations exploit the idea that truth-sets in Kripke-style semantics can be directly encoded as predicates (i.e. sets) in HOL. Modal formulas φ are thus identified with their corresponding truth sets $\varphi_{w\to o}$ of predicate type $w \to o$. In a similar vein, first-order predicates (i.e. operating on individuals of type e) have type $e\to w\to o$.

The semantic embeddings approach gives us two important benefits: (i) we can reuse existing automated theorem proving technology for HOL and apply it for automated reasoning in non-classical logics (e.g. free [11], modal [10], intuitionistic [9] and deontic [8] logics); and (ii) the logic of formalization becomes another degree of freedom in the development of a logical theory and thus can be fine-tuned dynamically by adding or removing axioms (which are encoded using HOL as a meta-language).

4.2 Example: Logical Analysis of a Metaphysical Argument

A first application of computational hermeneutics for the logical analysis of arguments has been presented in [35] (with its corresponding Isabelle/HOL sources available in [34]). In that work, a modal variant of the ontological argument for the existence of God, introduced in natural language by the philosopher E. J. Lowe [45], has been iteratively analyzed using our computational-

hermeneutic approach in a human-in-the-loop fashion and, as a result, a 'most' adequate formalization has been found. In each series of iterations (seven in total) Lowe's argument has been formally reconstructed in Isabelle/HOL employing varying sets of premises and logics, and the partial results have been compiled and presented each time as a new variant of the original argument (starting from Fig. 6 and ending with Fig. 7). In this fashion, Lowe's argument, as well as our understanding of it, gradually evolves as we experiment with different combinations of definitions, premises and logics for formalization.

Our computational-hermeneutic process starts with an initial, 'bootstrapped' formalization. As expected, such a first formalization is often non-well-formed and in most cases logically non-valid; it is generated by manually encoding in an interactive proof assistant (e.g. Isabelle) the author's natural-language formulation using our logically-educated intuition (and with some help from semantic parsers). After this first step, an interpretive give-and-take process is set in motion, and after a few iterations our candidate formalization starts taking shape, as shown in Fig. 6 where we have validated some of the argument's partial conclusions, though not the main one (where we get a countermodel from model finder Nitpick [15] in line 51). During this process we apply our human ingenuity to the abductive task of coming up with new formalization hypotheses (i.e. adding, removing and replacing axioms or definitions) and harness the power of Isabelle's automated reasoning tools to assess their adequacy.

Following the *principle of charity*, our tentative formalizations (logical theories) should become logically valid and meet some inferential (and sometimes syntactical) criteria to our satisfaction. In our case study, this has happened after the seventh iteration series, as shown in Fig. 7, where the now simplified argument[18] is not only logically valid (this happened already after the second iteration series), but also exhibits other desirable features of a more philosophical nature (as discussed in [35]); we speak here of arriving at a state of reflective equilibrium (e.g. following [20,50]).

In Fig. 7 we can also see how our set of axioms has become split into argument's premises (lines 26–30) and meaning postulates for its featuring concepts (lines 18–23). Here is also shown (line 23) how one of these postulates corresponds to a (meta-logical) axiom constraining the accessibility relation R, thus further framing the logic of formalization (modal logic KB). Note that the (meta-logical) definitions shown in Fig. 5, corresponding to the semantical embedding of modal logic K in Isabelle/HOL, can also be seen as further meaning postulates. This embedding has, indeed, also evolved during the interpretive process. For instance, we had to add at some point the formulas that embed restricted (so-called "actualist") first-order quantifiers (Fig. 5, lines 30–34), in order to adequately represent the argument (by not trivializing it, as explained in [35]). As a result of this computational-hermeneutic process, we can see how the meanings of some (rather obscure) metaphysical concepts, like "necessary existence",

[18] In [35] we have produced and discussed other more (or less) complex valid formalizations of this argument before finally settling with the one shown in Fig. 7.

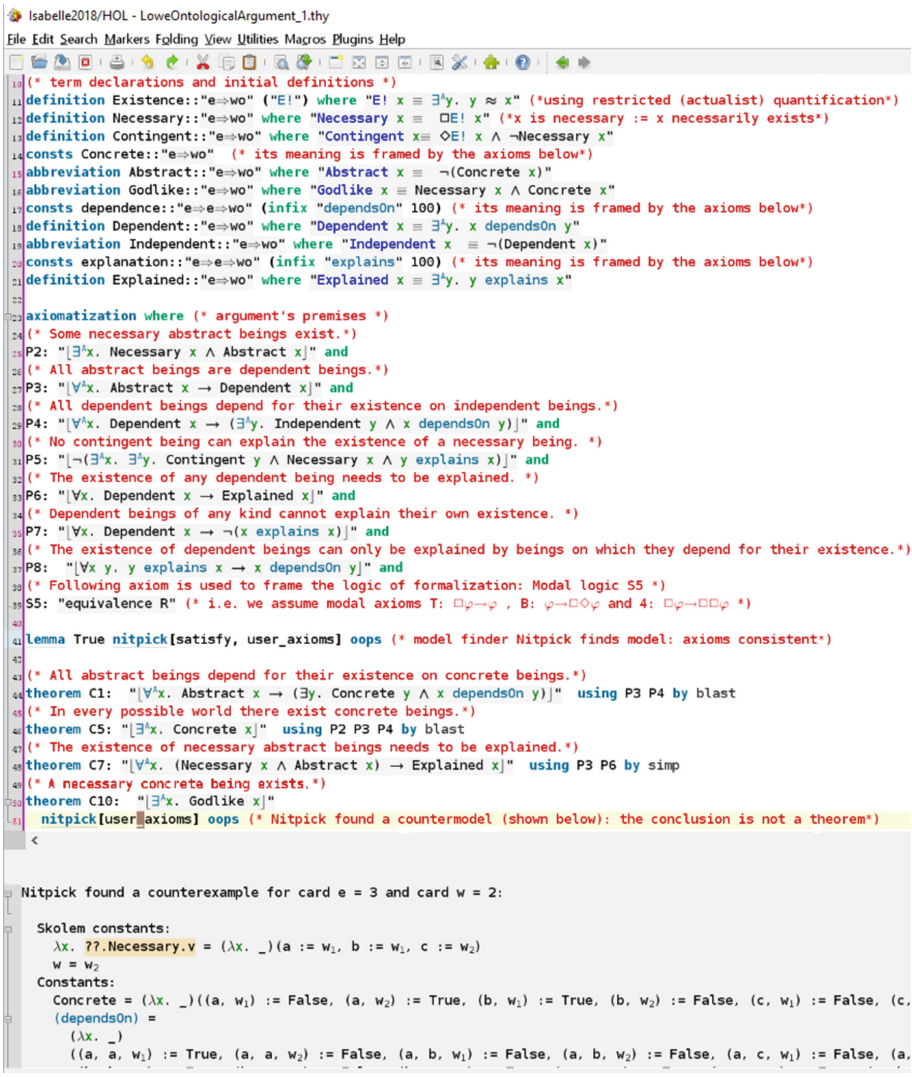


Fig. 6. A (non-valid) formalization of Lowe's argument resulting after the first iteration series. We can see that the model finder Nitpick [15] has found a countermodel.

"contingency", "abstractness/concreteness", "metaphysical explanation" and "dependence", have become explicitated in the form of logical formulas.

In follow-up work [36], we have extended the approach to logical analysis previously sketched in [35] by augmenting it with methods as developed in argumentation theory. The idea is now to consider the dialectic role an argument plays in some larger area of discourse (represented as a network of arguments).

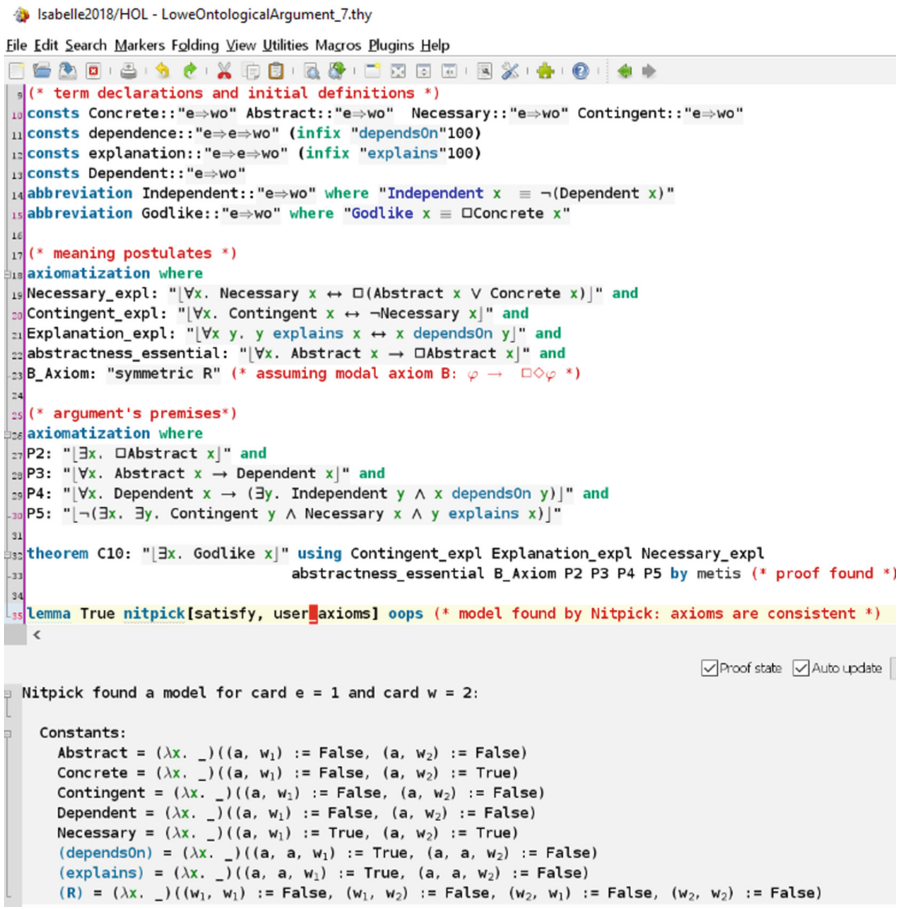


Fig. 7. Result after the last iteration series: a satisfying formalization. Notice the model generated by Nitpick for the set of axioms (indicating their consistency).

We have analyzed the contemporary debate around different variants of Gödel's ontological argument obtaining some interesting conceptual insights.

4.3 Technological Feasibility and Implementation Approaches

In this last section we want to briefly discuss some design ideas and challenges for a software system aimed at automating our computational-hermeneutic approach. These ideas are currently under development. By sketching a preliminary technological landscape, we argue for the technological feasibility of highly-automated computational hermeneutics in practice. In particular, we want to highlight some of the challenges involved. In Fig. 8, the main component of the system is the so-called "hermeneutic engine" (which will be expanded below in

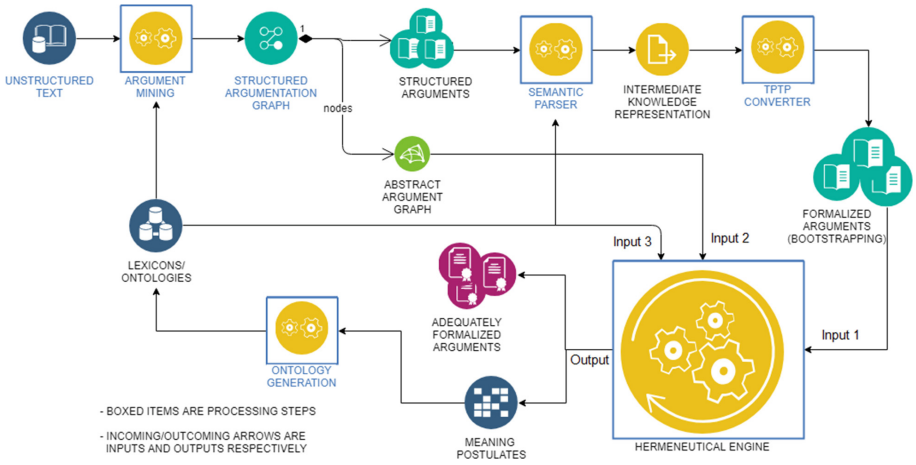


Fig. 8. The encompassing technological landscape.

Fig. 9). For the time being let us see it as a black box and discuss its inputs and outputs.

- **Input 1:** A collection of formalized arguments, i.e. sets of (labeled) formulas. We bootstrap the interpretive process with some initial rough and ready formalizations encoded in a special format for use with automated reasoning tools. We consider in particular the TPTP syntax [55] as it is well supported in most higher-order theorem provers. In order to arrive at these collections of (bootstrapped) formulas from our initial natural-language arguments, we rely on argumentation mining technology [21,44] and semantic parsers (e.g. the well-known *Boxer* system [17]). Depending on the technological choices, the output format of such tools will most probably be some kind of first-order representation, e.g. Discourse Representation Structures (DRS) (as in *Boxer*). We rely on existing tools supporting the conversion between those first-order representations and TPTP syntax.
- **Input 2:** Abstract argument network, i.e. a graph whose nodes are the labels of the formalized arguments provided in *Input 1* and whose edges correspond to attack or support relations. This kind of structures usually constitute the output of argumentation mining software [21,44].
- **Input 3:** Lexicons and ontologies play an important role in current approaches to semantic parsing and argumentation mining (e.g. the *Boxer* system has been extended with entity linking to external ontologies [1,38]). Furthermore, existing ontologies can serve as sources during the abductive step of coming up with candidate meaning postulates.
- **Output:** A collection of (improved) argument formalizations, encoded using appropriate expressive logics. Importantly, these formalizations have to be logically correct (in contrast to the initial, bootstrapped ones) and are not restricted to the kind of less-expressive first-order outputs of current semantic

parsing technology. During the process, we maintain a set of sentences common to most formalized arguments and scoring well according to our (mostly inferential) adequacy criteria. These sentences are then labeled as meaning postulates. (Recall from Sect. 2.1 that the heuristic to differentiate between meaning and empirical postulates is yet another degree of freedom, which is to be determined on the basis of pragmatic considerations.) In Sect. 2.1 we saw that an ontology (in computer science) can be thought of as a collection of meaning postulates. Thus, the output of our interpretive process can serve to create new domain-specific ontologies or extend existing ones.

Having discussed the inputs and outputs of our core component, we proceed to roughly outlining its operation in Fig. 9.

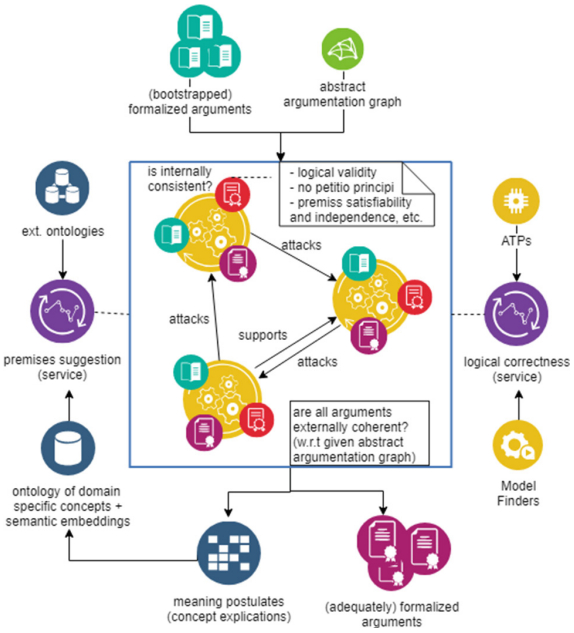


Fig. 9. Inside the "hermeneutic engine".

In a nutshell, the engine adds, removes and modifies the arguments' formalized sentences (*Input 1*) so that each argument in the network becomes logically correct, while still fulfilling its intended dialectic role in the network (*Input 2*). Formalized sentences are continuously rated (and picked out) using the adequacy criteria described above. Computational hermeneutics is thus an iterative and two-layered approach. In one layer we apply a formalization procedure to each individual argument in parallel, which consists in temporarily fixing truth-values and inferential relations among its sentences, and then, after choosing a logic for formalization (i.e. the set of axioms constituting the respective semantical

embedding in HOL), working back and forth on the formalization of its premises and conclusions while getting automatic feedback about the adequacy of the current formalization candidate. An adequate formalization is one which renders the argument as logically correct and scores high in some additional criteria (syntactic and inferential). In the next layer we operate at the network level, we continuously choose among the many different combinations of adequate argument formalizations those which honor their intended dialectic roles, i.e. those which render the arguments as successfully attacking or supporting other arguments in the network as intended. Importantly, we work at these two layers in parallel (or by continuously switching between them). In this fashion, by engaging in a methodical 'trial-and-error' interaction with the computer, we work our way towards a proper interpretation of a piece of argumentative discourse by circular movements between its parts and the whole. As shown in Fig. 9, the output (and usefulness) of our process consists in generating adequate, logically correct formalizations of natural-language arguments, while at the same time articulating logical formulas serving as meaning postulates for those concepts framed by the argumentative discourse in question. These formulas, following Carnap [23], can be seen as conceptual *explications*, thereby fulfilling the task of making the discourse's implicit conceptualization explicit.

5 Conclusion

Guarino et al. [42,43] saw in their notion of a (formal) conceptualization a touchstone for evaluating how good a given logical theory (i.e. an ontology) models or approximates our conception of reality. As a result of this correspondence with the 'real world', inferences drawn using the theory will lead to intuitively correct conclusions. Drawing on Guarino's exposition and terminology, we have inverted the situation and argued for the thesis that a conceptualization first becomes articulated by explicitly stating a theory. We can even put it more dramatically and say that a conceptualization first comes into existence by its being disclosed through an *adequate* logical theory (or more specifically: its models).[19] For the sake of our self-conscious analysis of socio-linguistic practices, we can at any point during an interpretive endeavor equate a conceptualization with the collection of models of the logical theory under consideration, provided that this theory satisfies certain adequacy criteria. This has lead us to discuss what makes an interpretive, logical theory "adequate" in this sense. Drawing upon a holistic view of meaning and logical form, we have approached this issue by considering intersubjective argumentative practices (instead of 'objective reality') as our touchstone. That is, the inferences licensed by our adequate theory are to correspond in a way with the inferences (arguments) endorsed by a community of speakers (and always in the context of some specific discourse). We briefly presented Donald Davidson's account of radical interpretation [27] together with some inferential adequacy criteria of formalization recently

[19] Such a reading would be in tune with strong conceptions of existence drawing on the Quinean slogan "no entity without identity".

proposed in the philosophical literature [50] and discussed how they relate to our approach. In the last section we have presented an exemplary application of (human-in-the-loop) computational hermeneutics and sketched some technical implementation approaches towards its further automation. We showed how current technologies in computational linguistics and automated reasoning enable us to automate many of the tasks involved in our (currently human-supported) computational-hermeneutic approach. Other, difficult to automate tasks, in particular those concerned with abductive reasoning (in higher-order and/or non-classical logics), pose interesting challenges that we will be addressing in future work. There are ongoing efforts to frame our approach in terms of a combinatorial optimization problem. Given the underdetermined, inexact nature of the problem of logical analysis (and the undecidability of theorem proving in higher-order logics), we do not expect exact algorithms to be found and focus primarily on (highly-parallelizable) metaheuristics. The integration of machine learning techniques into our approach, in particular regarding the abductive task of suggesting missing (implicit) argument premises, is also being contemplated.

References

1. Basile V, Cabrio E, Schon C (2016) KNEWS: using logical and lexical semantics to extract knowledge from natural language. In: Proceedings of the European conference on artificial intelligence (ECAI) 2016 conference
2. Baumberger C, Brun G (2016) Dimensions of objectual understanding. In: Explaining understanding. New perspectives from epistemology and philosophy of science, pp 165–189
3. Baumgartner M, Lampert T (2008) Adequate formalization. Synthese 164(1):93–115
4. Benzmüller C (2019) Universal (meta-)logical reasoning: recent successes. Sci Comput Program 172:48–62
5. Benzmüller C, Andrews P (2019) Church's type theory. In: Zalta EN (eds.) The stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University, summer 2019 edition
6. Benzmüller C, Brown C, Kohlhase M (2004) Higher-order semantics and extensionality. J Symbolic Logic 69(4):1027–1088
7. Benzmüller C, Fuenmayor D (2018) Can computers help to sharpen our understanding of ontological arguments? In: Gosh S, Uppalari R, Rao KV, Agarwal V, Sharma S (eds) Mathematics and Reality. Proceedings of the 11th All India Students' Conference on Science & Spiritual Quest (AISSQ). The Bhaktivedanta Institute, Kolkata, pp 195–226
8. Benzmüller C, Parent X, van der Torre L (2018) A deontic logic reasoning infrastructure. In: Manea F, Miller RG, Nowotka D (eds) Proceedings of the 14th conference on computability in Europe (CiE), LNCS, vol 10936, pp 60–69. Springer
9. Benzmüller C, Paulson L (2010) Multimodal and intuitionistic logics in simple type theory. Logic J IGPL 18(6):881–892
10. Benzmüller C, Paulson L (2013) Quantified multimodal logics in simple type theory. Log Univers 7(1):7–20 (Special Issue on Multimodal Logics)
11. Benzmüller C, Scott DS (2019) Automating free logic in HOL, with an experimental application in category theory. J Autom Reasoning

12. Benzmüller C, Weber L, Woltzenlogel Paleo B (2017) Computer-assisted analysis of the Anderson-Hájek controversy. Log Univers 11(1):139–151
13. Benzmüller C, Woltzenlogel Paleo B (2016) The inconsistency in Gödel's ontological argument: a success story for AI in metaphysics. In: Kambhampati S (eds) IJCAI 2016, vol 1–3, pp 936–942. AAAI Press
14. Besnard P, Hunter A (2008) Elements of argumentation. MIT Press (2008)
15. Blanchette JC, Nipkow T (2010) Nitpick: a counterexample generator for higher-order logic based on a relational model finder. In: Kaufmann M, Paulson LC (eds) ITP 2010, vol 6172. LNCS. Springer, Heidelberg, pp 131–146. https://doi.org/10.1007/978-3-642-14052-5_11
16. Blau U (1978) Die dreiwertige Logik der Sprache: ihre Syntax, Semantik und Anwendung in der Sprachanalyse. Walter de Gruyter
17. Bos J (2008) Wide-coverage semantic analysis with boxer. In: Bos J, Delmonte R (eds.) Semantics in text processing, STEP, 2008 Conference proceedings, Venice, Italy, 22–24 September 2008. Association for Computational Linguistics (2008). https://dblp.org/rec/bib/conf/acl-step/Bos08a
18. Brandom RB (1994) Making it explicit: reasoning, representing, and discursive commitment. Harvard University Press
19. Brun G (2004) Die richtige Formel: Philosophische Probleme der logischenFormalisierung. Walter de Gruyter
20. Brun G (2017) Conceptual re-engineering: from explication to reflective equilibrium. Synthese, pp 1–30
21. Budzynska K, Villata S (2018) Processing natural language argumentation. In: Baroni P, Gabbay D, Giacomin M, van der Torre L (eds) Handbook of formal argumentation, pp 577–627. Springer
22. Burgess A, Plunkett D (2013) Conceptual ethics I & II. Philos Compass 8(12):1091–1110
23. Carnap R (1947) Meaning and necessity: a study in semantics and modal logic. University of Chicago Press
24. Carnap R (1952) Meaning postulates. Philos Stud 3(5):65–73
25. Davidson D (1994) Radical interpretation interpreted. Philos Persp 8:121–128
26. Davidson D (2001) Essays on actions and events: philosophical essays, vol 1. Oxford University Press on Demand
27. Davidson D (2001) Inquiries into truth and interpretation: philosophical essays, vol 2. Oxford University Press
28. Dowty DR, Wall R, Peters S (2012) Introduction to montague semantics, vol 11. Springer
29. Doyle J (1992) Reason maintenance and belief revision: foundations vs. coherence theories. Belief Revision 29:29–51
30. Dung PM, Kowalski RA, Toni F (2009) Assumption-based argumentation. In: Argumentation in artificial intelligence, pp 199–218. Springer
31. Elgin C (1999) Considered judgment. Princeton University Press
32. Epstein RL (1994) The semantic foundations of logic: predicate logic, vol 2. Oxford University Press
33. Fuenmayor D, Benzmüller C (2017) Automating emendations of the ontological argument in intensional higher-order modal logic. In: Kern-Isberner G, Fürnkranz J, Thimm M (eds) KI 2017: Advances in artificial intelligence. LNAI, vol 10505, pp 114–127. Springer
34. Fuenmayor D, Benzmüller C (2017) Computer-assisted reconstruction and assessment of E. J. Lowe's modal ontological argument. Archive of formal proofs, September 2017. http://isa-afp.org/entries/Lowe_Ontological_Argument.html, Formal proof development

35. Fuenmayor D, Benzmüller C (2018) A case study on computational hermeneutics: E. J. Lowe's modal ontological argument. J Appl Logics (IfCoLoG J Logics Appl) 5(7):1567–1603 (special issue on Formal Approaches to the Ontological Argument)
36. Fuenmayor D, Benzmüller C (2019) Computational hermeneutics: an integrated approach for the logical analysis of natural-language arguments. In: Liao B, Agotnes T, Wang YN (eds) Dynamics, uncertainty and reasoning: the second Chinese conference on logic and argumentation
37. Gadamer HG (1960) Gesammelte Werke, Bd. 1, Hermeneutik I: Wahrheit und Methode. J.C.B. Mohr (Paul Siebeck)
38. Gangemi A, Presutti V, Recupero DR, Nuzzolese AG, Draicchio F, Mongiovì M (2017) Semantic web machine reading with FRED. Seman Web 8(6):873–893
39. Genesereth MR, Nilsson NJ (1987) Logical foundations of artificial intelligence. Morgan Kaufmann
40. Goodman N (1983) Fact, fiction, and forecast. Harvard University Press
41. Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquisition 5(2):199–220
42. Guarino N, Giaretta P (1995) Ontologies and knowledge bases towards a terminological clarification. In: Towards very large knowledge bases: knowledge building & knowledge sharing, vol 25, no 32, pp 307–317
43. Guarino N, Oberle D, Staab S (2009) What is an ontology? In: Handbook on ontologies, pp 1–17. Springer
44. Lippi M, Torroni P (2016) Argumentation mining: state of the art and emerging trends. ACM Trans Internet Technol (TOIT) 16(2):10
45. Lowe EJ (2013) A modal version of the ontological argument. In: Moreland JP, Sweis KA, Meister CV (eds) Debating Christian theism, chapter 4, pp 61–71. Oxford University Press
46. Nipkow T, Paulson L, Wenzel M. (2002) Isabelle/HOL: a proof assistant for higher-order logic. LNCS, lecture notes in computer science, vol 2283. Springer
47. Novaes CD (2018) Carnapian explication and ameliorative analysis: a systematic comparison. Synthese, pp 1–24
48. Peregrin J (2014) Inferentialism: why rules matter. Springer
49. Peregrin J, Svoboda V (2013) Criteria for logical formalization. Synthese 190(14):2897–2924
50. Peregrin J, Svoboda V (2017) Reflective equilibrium and the principles of logical analysis: understanding the laws of logic. Routledge studies in contemporary philosophy. Taylor and Francis
51. Quine WVO (1960) Word and object. MIT Press
52. Rawls J (2009) A theory of justice. Harvard University Press
53. Sainsbury M (1991) Logical forms: an introduction to philosophical logic. Blackwell Publishers
54. Studer R, Benjamins VR, Fensel D (1998) Knowledge engineering: principles and methods. Data Knowl Eng 25(1–2):161–197
55. Sutcliffe G (2017) The TPTP problem library and associated infrastructure. From CNF to TH0, TPTP v6.4.0. J Autom Reason 59(4):483–502
56. Tarski A (1956) The concept of truth in formalized languages. In: Logic, semantics, metamathematics, vol 2, pp 152–278
57. Uschold M (1996) Building ontologies: towards a unified methodology. In: Proceedings of 16th annual conference of the British computer society specialists group on expert systems. Citeseer



The Context-Priming of Conceptual Knowledge: RPEC and SPEC Models

Bideng Xiang[1(✉)] and Ping Li[2]

[1] School of Electronic Commerce, Jiujiang University, Jiujiang 332005, China
xiangbideng@126.com
[2] Department of Philosophy, Sun Yat-sen University, Guangzhou 510275, China
hsslip@mail.sysu.edu.cn

Abstract. There have been three viewpoints about the role of contexts played in the priming processes of conceptual knowledge: Some conceptual psychologists claim that priming effects of concepts are independent of any specific contexts; others contend that the core of concepts is activated automatically though the remainder is context-dependent; and recent studies show that all the properties contained in concepts are primed in specific contexts. This paper will propose two context-dependent models of priming conceptual knowledge, *Referent-Priming-Effects in Context* (RPEC) and *Semantics-Priming-Effects in Context* (SPEC), in order for a defense of the last view, namely *the Context-Priming Hypothesis of Conceptual Knowledge*. The first model indicates that there must be spatial and temporal connections between the priming and the primed items while the second reveals that there must be particular semantic connections between the priming and the primed. Thus the two models present two fundamental context-dependent modes of priming conceptual knowledge in spite of a variety of context-dependence.

1 Introduction

With their respective concerns, philosophers of science have dealt with concepts that are considered as a body of knowledge about the properties of particular categories in psychology for over two decades (for example, Machery 2009; Nersessian 1999a; 1999b; 2002; Andersen et al. 1996; Chen et al. 1998). Likewise, the present paper will discuss a different issue related to the priming of conceptual knowledge. The so-called priming effect is a kind of mental phenomena in which the concepts processed previously in a cognitive process would activate other concepts or speed up the cognitive processes related to them. In such priming processes the concepts manipulated first are called priming-items (or primes and cues); and the activated is called primed-items (or targets). For instance, the categorization of green emerald would accelerate the categorization of bamboo if the color *green* is highlighted.

As a complex body of knowledge, relevant concepts stored in long-term memory would be primed or activated only when a specific cognitive task is executed. The concepts of metal and wood will be activated, for example, if it's needed

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 470–481, 2019.
https://doi.org/10.1007/978-3-030-32722-4_26

to judge whether an object is metal or wood. In respect to a certain concept, however, whether the whole or some parts of conceptual knowledge would be primed in a particular cognitive task? And which part/parts would be primed on earth if some parts would be primed? Additionally, whether to prime conceptual knowledge is context-independent or context-dependent? All these questions have always been along with controversies.

Barsalou (1982) proposes that concepts contain two types of properties, context-independent and context-dependent properties. The former are activated by the words for their respective concepts on all occasions, and the latter are not activated by the corresponding words independent of context. And some experimental studies show that high-dominant properties (e.g., barking is a high-dominant property of dogs) "function as core, or invariant, aspects of meaning and that initial semantic access is context independent" (Whitney et al. 1985, p. 126). Moreover, Machery (2009) claims that conceptual knowledge is used by default, namely all knowledge of a specific concept would be activated automatically at once when it used in a certain cognitive process.

In contrast to these views, it's claimed that there are no such important properties that constitute the core of a concept and would be activated automatically. For example, it's contended that all the knowledge in a particular concept is context-sensitive. (Lebois et al. 2015) The experiments conducted by Pecher and Raaijmakers suggest that "The responses to a word (e.g. *bread*) are faster and more accurate if the target word is presented in the context of an associated word, the prime (e.g. *butter*), than if it is presented in the context of an unrelated prime (e.g. *chair*)." (Pecher and Raaijmakers 1999, p. 593) Other experimental researches also show that "Only after activation tasks that focused on perceptual features was priming for perceptually related word pairs found in pronunciation." (Pecher et al. 1998, p. 401) As far as abstract concepts are concerned, they can activate one another. As McRae et al. (2018, p. 518) write: Abstract concepts "are not devoid of perceptual information because knowledge of real-world situations is an important component of learning and using many abstract concepts."

Against the automatic-priming principle, this paper will present two context-dependent models of priming conceptual knowledge in Sect. 2, based on the recent studies of context-priming effects, and then make arguments for the context-priming hypothesis of conceptual knowledge in Sects. 3 and 4.

2 Two Context-Dependent Models of Priming Conceptual Knowledge

Generally, context-priming effects can generate in two ways. In one way, the prime(s) and target(s) share a common physical context, and there are particular spatial and temporal relations between them. The context-priming effects produced in this way are called *Referent-Priming-Effects in Context* (RPEC) because those priming effects are just relevant to the referents of priming and primed concept(s), rather than their semantics. In another way, the prime(s)

and target(s) share some similar or same properties, and there are particular semantic relations between them. The context-priming effects produced in this way are called *Semantics-Priming-Effects in Context* (SPEC) because such priming effects are just relevant to the semantics of priming and primed concept(s), rather than their referents.

2.1 Referent-Priming-Effects in Context (RPEC)

Van Dantzig et al. (2011, p. 1164) suggest that "a context can ... be defined as a coherent collection of objects that co-occur within a certain window of time and space." For example, all kinds of kitchenware and the particular spatial and temporal connections between them constitute a functional context – kitchen. They also contend that there exist co-conceptualizing effects between a context and its elements. In other words, a context and its elements can be conceptualized simultaneously under the common spatial and temporal conditions. The co-conceptualization of context elements is a cognitive process in which an agent develops gradually conceptual representations of context elements by experiencing similar or same contexts repeatedly. On the other hand, the co-conceptualization of context is a cognitive process in which an agent develops gradually conceptual representations of the context itself and the spatial-temporal connections among context elements at the same time when they are conceptualized.

According to these descriptions of context and co-conceptualizing effects, context-priming effects are such a kind of mental phenomena in which the concepts of some elements in a specific context (such as a refrigerator in kitchen) can activate the concepts of other kitchen wares (such as cupboard, stove, and so on). And in such context-priming processes priming-items and primed-items contain the concepts of any objects in the context, including the context itself (Van Dantzig et al. 2011).

Four significant conclusions can be made in terms of the above characterization of context, co-conceptualization and context-priming effects: (1) priming-items and primed-items in context-priming processes share a common context; (2) context elements contain not only physical objects, but also the particular spatial-temporal connections between them; (3) a conceptualization of context is based on the conceptualization of its elements and the particular spatial-temporal relations between them – that is, the former is the function of the latter; and (4) the concept of context and its elements constitute a systematic and hierarchical structure.

Furthermore, the particular spatial-temporal connections between context elements are necessary for co-conceptualization between context elements while a co-conceptualization between context elements is a cognitive base of context-priming effects between the concepts of both context elements and context itself. For instance, refrigerator, cupboard and other kitchen wares can be co-conceptualized because in a specific time they are located at a specific position in a common context (kitchen). And then the concepts of refrigerator and other kitchen wares can activate each other when one of them is used in some cognitive tasks.

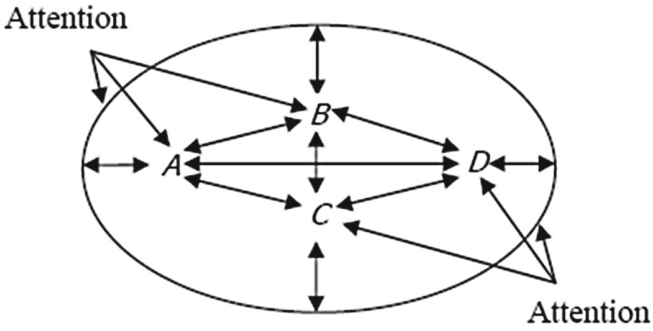


Fig. 1. Referent-priming effects

Since there exist some particular spatial-temporal links between context elements that stay in a common context, priming effects would generate between the priming and primed concepts as long as they could respectively refer to their objects successfully. Such a kind of context-priming effects is called *Referent-Priming-Effects in Context* (RPEC) in this paper, which appeals to the referring function of concepts to produce priming effects. Figure 1 illustrates the model of RPEC, in which the oval denotes the whole context (e.g., kitchen), the capital letters *A*, *B*, *C* and *D* denote the physical elements of context (e.g., refrigerator, cupboard, and so on), and the two-headed arrows denote mutual priming effects between the context and its elements. Besides, it's necessary that the element(s) or the context itself is/are highlighted or focused on. That is, only after some elements or the context itself was paid attention to in a particular cognitive task, could these elements prime each other.

2.2 Semantics-Priming-Effects in Context (SPEC)

Existing experimental evidence or findings support the hypothesis that there exist priming effects between the concepts that have similar properties (e.g., Pecher et al. 1998; Yee et al. 2012). The so-called similarity or overlap of features between priming and primed items means that there are some important properties shared by both priming concepts and primed ones. Given the property *white* used to be a diagnostic feature common to cottons and clouds in experiments, for instance, some experimental findings suggest that retrieving conceptual knowledge about object-associated colors and color perception share neural bases (Simmons et al. 2007; Hsu et al. 2011). This perhaps implies that the reason why the concepts with similar properties can activate mutually is that the information about similar features of different concepts is manipulated by the same or similar neural mechanisms.

In fact, Pecher and Yee et al.'s findings suggest a different foundational context-dependent mode of priming conceptual knowledge. Besides the mode mentioned in the previous section, *the Referent-Priming-Effects in Context*, which are based on the spatial-temporal relations between priming concepts and

primed ones, there is another: Primes and targets can also prime one another if there are semantic connections between both of them. In other words, two concepts without similar or same properties would not generate such a kind of priming effects based on semantic links between them since there is no any semantic connections linking them together. Indeed, there is evidence that "priming in word association depends largely on the storage of information relating the cue and the target" (Zeelenberg et al. 2003). We call this kind of context-priming effects as *the Semantics-Priming-Effects in Context* (SPEC) instantiated by Fig. 2.

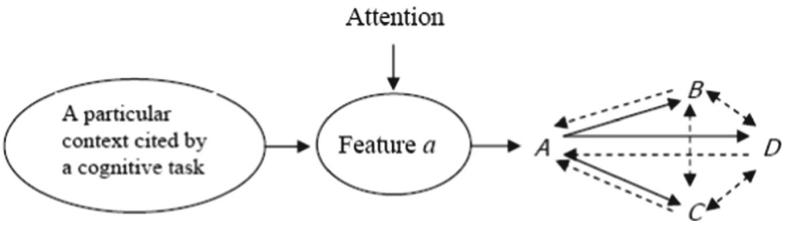


Fig. 2. Semantics-priming effects

Figure 2 indicates that the property a of the concept A is highlighted in a specific context or cognitive task, in which the capitals A, B, C and D denote four different concepts that have similar properties, the solid arrows denote direct priming processes, and the dotted arrows denote indirect priming processes. After judging the color of cotton, for example, the property *white* of the concept *cotton* was paid attention to and became outstanding. Thus the concept *cotton* and other concepts sharing the same property *white*, such as *cloud*, *snow*, and *swan*, would activate each other. Among these concepts, the concept *cotton*, activated first, would prime other concepts directly, which could activate in turn each other indirectly.

3 Constraints and Interference of Context-Priming Effects

Although the two models characterized above have respectively proposed the necessary conditions for context-priming effects, priming effects would not generate because of a lack of other conditions – e.g., attention and neural bases as constraints – and might be affected by many interference factors such as beliefs, preferences and emotions, which can be explained within the conceptual framework of the context-priming hypothesis of conceptual knowledge.

3.1 Attention Paid to the Concepts of Some Elements of Context or to Some Features Shared by the Priming and Primed Items

Pecher et al. (1998) find out that the word pairs whose concepts have perceptually similar features would not prime each other automatically and that they

would prime mutually only after some perceptual features are activated in relevant cognitive tasks. For example, the word pairs such as *coin-medal* would prime each other after judging the shape of clock since they refer to objects with similar or same shape. Yee et al. (2012) also show that only after subjects' attention being transferred to color by a Stroop task (viz. judging the color of the letters composing a word which denotes a particular color) would reading the word referring a specific object prime the concepts of other objects that hold the same diagnostic feature *color* with it. When the color of cotton is noticed, for instance, reading the word *cotton* would prime the word *cloud*, since in general cotton and cloud have the same diagnostic color (white).

Still there is evidence that the features contained in a concept would not be activated automatically, only the features highlighted in specific contexts would be activated and prime other concepts that have those similar features (Lebois et al. 2015). For instance, only after it being seen that a clock has fallen down and broken a cup, could the weight and location etc be activated while other perceptual properties, such as shape and color, would not be activated; and also it would not prime the concepts of coin and medal though they have the same shape with the clock.

Before some elements of a particular context or the context itself haven't been noticed, as far as the referent-priming effects are concerned, maybe any one of the elements and the context itself would not prime each other.

3.2 Particular Neural Systems Needed to Organize the Conceptual Knowledge on Which Context-Priming Effects Rely

The concepts developed in the process of co-conceptualization between a context and its elements contain visual, aural, and other kinds of perceptual information. In order to encode, restore, deliver and retrieve such kinds of conceptual information, especially for storing and retrieving systematically in the light of the spatial-temporal relations developed during co-conceptualization, some neural systems with special functions would be indispensable. For instance, people could remember the shape and site of all kinds of kitchenware and even the taste of some foods after visiting a friend's kitchen. When they talk about that visit someday, it would be needed to retrieve all the information acquired during that visit. In such a situation a particular neural system is necessary to reserve and retrieve the information of all kinds of kitchenware and foods chronologically. Besides the conceptual information concerning context-referent priming effects, the conceptual information related to context-semantics priming effects needs also to be organized by particular nervous systems. For example, the information about similar properties of priming and primed items should be reserved and retrieved systematically; therefore a particular neural system is needed. As an illustration, some experimental findings suggest that retrieving conceptual knowledge about object-associated color properties and color perception share neural bases (Simmons et al. 2007; Hsu et al. 2011).

Meyer and Damasio (2009) claim that there exist convergence-divergence zones (CDZs) in the brain that function as association areas to register the

temporal co-occurrences between patterns of activation in lower-level areas. As they write: "The architecture is constituted by two crucial elements: (i) neuron ensembles in early sensory and motor cortices, which represent separate knowledge fragments about a given object; and (ii) neuron ensembles located downstream from the former in association cortices, which operate as convergence-divergence zones (CDZs ...)." (Meyer and Damasio, 2009, p. 376) During interaction with an object, patterns of activation arise across different sensorimotor areas, coding for various aspects of the interaction. The temporally coincident activity at separate sites are bound by a CDZ. In order to recall a particular event, the CDZ (partially) reconstructs the original pattern of activation in the sensorimotor sites through divergent back projections. Meyer and Damasio argue that multiple convergence zones can be organized in a hierarchical manner, each representing input at different levels of abstractness. That is, there are convergence zones of different orders. For example, some zones bind features into entities, and others involve entities into events or sets of events; but all such zones register the combinations of components in terms of coincidence or sequence in space and time.

3.3 Interference Factors of Context-Priming Effects

As Sect. 3.1 mentioned, there must be some similar or same properties shared by the priming and primed concepts in context-semantics priming processes, which must be highlighted in relevant cognitive tasks. Hypothetically, some elements of a particular context or the context itself should be noticed first in context-referent priming processes. However, whether (or to what extent) those properties would be focused on is interfered by many mental factors.

First, specific beliefs (even false beliefs) would strengthen context-semantics priming effects. For instance, reading the word *crow* would prime the concepts *cotton*, *snow*, and so on if the subjects believe that all the crows are white. On the contrary, reading the same word *crow* would not prime the concepts *cotton*, *snow*, and etc if the subjects do know and believe that all the crows are black.

Second, specific preferences would strengthen context-semantics priming effects. If the subjects favor the color *red*, say, reading the word *flamingo* would prime the concept *morning glow* more easily than those who dislike the color *red*.

Third, specific emotions would affect context-semantics priming effects. Obviously happiness and fear would affect priming processes differently.

On the other hand, context-referent priming processes should also be interfered by these factors in different ways. When the subjects favor apple, for instance, it would be easier to prime the concepts of other fruits in a store than if they don't favor apple. Of course, there should be other subjective factors disturbing context-priming effects although the underlying mechanisms are still unclear.

Taken together, there are still many aspects relevant to context-priming effects keep unclear, such as other constraints and interfering factors, especially the corresponding neural bases and the mechanism by which attention plays role in the priming courses.

4 The Advantages of the Context-Priming Hypothesis

It is difficult for the automatic-activation or default-using principle to deal with the issues stated in Sect. 1. But the context-priming hypothesis of conceptual knowledge seems better in doing so. In this sense, the hypothesis has its advantages. The following paragraphs will present its answers to those questions.

4.1 Context Can Determine Which Kind/kinds of Concepts Would Be Primed

According to the default-using principle proposed by Machery (2009), conceptual knowledge is stored in long-term memory and used by default. That is, all the conceptual knowledge is activated automatically. In his argument for the principle Machery provides a linguistic case, in which people tend to give a positive response when they hear or read a sentence *a cheetah can outrun a man*. Machery contends that people would not give a positive answer to the sentence *a cheetah can outrun a man* if the concept of cheetah is not activated by default. So he takes this case as a strong support to his default-using tenet.

According to the context-referent priming model, however, the sentence *a cheetah can outrun a man* just cites a general context. As the elements of context, cheetah, man, the whole context itself, and the relations between all of them constitute a general context in which the concepts of cheetah and man are just their prototype concepts. When the sentence is changed to *a cheetah always outrun a man*, it is concerned with a bigger context that contains some special contexts besides the general one. In the case that a special context may include a young or sick cheetah and the strongest long-distance runner, people would not give a positive response again when they hear or read the changed sentence. This situation shows that the bigger context can activate the exemplar concepts of cheetah and man as well as their prototype concepts. So it can be concluded that context can determine which kind/kinds of concepts would be primed.

4.2 Context Can Determine Which Property/properties of a Particular Concept Would Be Activated

In the linguistic case mentioned in the previous section, when the sentence *a cheetah can outrun a man* activates the prototype concepts of cheetah and man by citing a general context, it can activate some particular properties (e.g. long legs, running fast, and etc) contained in those prototype concepts simultaneously. If the sentence is changed to another one *a cheetah can catch an antelope*, however, it might activate other particular features (e.g., having sharp tooth and claws). So it can be imagined that different sentences about cheetah can cite different contexts while different contexts can activate different kinds of concepts and different features of cheetah. That is to say, eventually context can determine which property/properties of a concept would be activated; and the different kinds and features of concepts would not be used by default across contexts.

4.3 Context Can Determine Whether the Concepts Themselves Would Be Activated

Moreover, context can determine the meaning of a sentence or whether the concepts themselves would be activated. Consider another linguistic case. When subjects hear or read the sentence *I don't want to be a sheep*, generally they don't understand the meaning of this sentence. Maybe they would ask what's wrong with the sheep. On those occasions, there would be no any conceptual knowledge of sheep activated, namely the whole concept of sheep wouldn't be activated. If subjects hear or read the sentence *I don't want to be a sheep and also a wolf*, however, they would really understand the true meaning of this sentence – that is, the sheep is too weak and the wolf is too aggressive. In conclusion, the single sentence *I don't want to be a sheep* can't activate the concept of sheep because of a lack of citing an effective context while the complex sentence *I don't want to be a sheep and also a wolf* can activate the concepts of both sheep and wolf due to its offering of an effective context.

4.4 Only the Context-Priming Processes Could Really Guarantee Cognitive Efficiency

In their connectionist model of learning conceptual knowledge (CONCAT), Van Dantzig et al. (2011) claim that a conceptual system is organized as a hierarchy with two levels: The first one is concepts of objects developed by extracting statistical regularities in feature co-occurrence of objects; and the second is the concept of context developed by extracting statistical regularities in objects co-occurrence of the whole context. Yeh and Barsalou (2006) contend that the correlations between entities and events that tend to occur in particular contexts more than others can constrain and facilitate relevant cognitive processes. That is, there is a need to retrieve specific information across all situations; and the cognitive system can use the current situations to aim at the specific concepts of contexts or objects and events in long-term memory. On the contrary, the cognitive system can use the current specific concepts of contexts or objects and events to aim at the situations. Since the cognitive system can focus on the most relevant context or the concepts of its objects and events directly in this way, the context-priming effects have efficient strengths. Conversely, the automatic-activation mechanism advocated by Machery and others cannot facilitate such cognitive processes.

5 Discussion and Conclusion

As a matter of fact, orthographic similarity effects and phonological similarity effects are two kinds of very interesting priming effects: orthographically similar words (e.g., *sand* and *wand*) can mutually activate; and phonologically similar words (e.g., *bait* and *gate*) can prime each other. Some experiments (e.g., Pecher et al. 2005) show that the orthographically similar words (called as *neighbors*)

can affect the performance in semantic-classification tasks. Specifically, subjects' performance is strengthened if the target (e.g., *cat*) belongs to the same category with the prime (e.g., *rat*); on the contrary, the performance is weakened if the target (e.g., *cat*) comes from the opposite class of the clue (e.g., *mat*).[1] Other experiments (e.g., Pecher et al. 2011) suggest that the performance differences between target words with phonologically coincident orthographic neighbors and those without phonologically coincident orthographic neighbors were larger when the orthographic neighbors were also semantically coincident (e.g., word pairs *book-hook* and *sand-wand*). At the same time, the performance differences between target words with semantically coincident orthographic neighbors and those without semantically coincident orthographic neighbors were larger when the orthographic neighbors were also phonologically coincident (e.g., word pairs *cat-rat* and *cat-mat*).

A little current literature deals with both orthographic similarity effects and phonological similarity effects as semantic priming ones. Regardless of the unclear cognitive bases or mechanisms underlying orthographic, phonological and semantic priming effects, however, it can be claimed that those effects really exist and can strengthen each other. That is, the three kinds of effects can overlap or resonate mutually. For example, an activation between the words *cat* and *rat* is the easiest because there exists a triple similarity – i.e. orthographic, semantic, and phonological similarity – between them. As far as the context-dependent priming effects of conceptual knowledge are concerned, orthographic similarity effects and phonological similarity effects could be attributed to context-semantics priming effects in the sense that they are dependent on previous cognitive processing of primes or when the orthography and phonology of a word were considered as the indispensable components of its concept. No matter whether they are treated by such a way or not, there really are some mental or neural links between the orthographically or phonologically similar word pairs. Otherwise, there wouldn't be any orthography similarity effects and phonology similarity effects. Anyway, they would not be against or contradict the context-semantics priming principle.

Nevertheless, it's not incompatible utterly between the context-priming and the automatic-activation principle. When the concept of a category is simple or contains very few features, in particular, the words labeling some particular categories can correlate the conceptual knowledge of those categories. In the case that children learn the concept of dog at the early stage, for instance, they could associate the feature *barking* when they hear the word *dog*; and at the same time, they could associate the concept of dog (e.g., an exemplar of their pet dogs) when they hear the barking voice. But this activation pattern of conceptual knowledge is so special that it's impossible to become both a general pattern and the dominant pattern of activating conceptual knowledge.

In conclusion, the context-priming hypothesis of conceptual knowledge has stronger explanatory power than the automatic-activation principle since

[1] According to the animate-inanimate category used in the experiments (Pecher et al. 2005), cat and rat belong to the same category *animate* while mat to *inanimate*.

the former is useful both for an explanation of the significant constraints and interfering factors of priming processes and for an answer to all the controversial questions mentioned in the introduction section. On the one hand, the automatic-activation or default-using principle ignores both the important constrains on priming effects and the basic factors affecting priming processes of conceptual knowledge, which constitute significant aspects of context-dependence. On the other hand, the context-priming hypothesis has the advantage of the automatic-activation or default-using principle in answering the relevant controversial questions. It is the former that provides definite answers: Context can determine which kind/kinds of concepts would be primed, which property/properties of a particular concept would be activated, and whether the concepts themselves would be activated; and only the context-priming processes could really guarantee cognitive efficiency. As a defense of the context-priming hypothesis, finally, this paper makes efforts to do two extended pieces of work: One is to build two context-dependant models of priming conceptual knowledge, *RPEC* and *SPEC*, as a generalization of recent experimental findings related to two fundamental context-dependent modes of priming conceptual knowledge; another is to reconcile orthographic similarity effects and phonological similarity effects with semantic priming ones for the purpose of eliminating a potential threat against the context-priming hypothesis.

Acknowledgements. The research work of this article was supported by a grant (18ZDA033) from the National Social Science Foundation of P.R. China.

References

Andersen H, Barker P, Chen X (1996) Kuhn's mature philosophy of science and cognitive psychology. Philos Psychol 9:347–363

Barsalou LW (1982) Context-independent and context-dependent information in concepts. Mem Cogn 10:82–93

Chen X, Andersen H, Barker P (1998) Kuhn's theory of scientific revolutions and cognitive psychology. Philos Psychol 11:5–28

Hsu NS, Kraemer DJM, Oliver RT, Schlichting ML, Thompson-Schill SL (2011) Color, context, and cognitive style: variations in color knowledge retrieval as a function of task and subject variables. J Cogn Neurosci 23:2544–2557

Lebois LAM, Wilson-Mendenhall CD, Barsalou LW (2015) Are automatic conceptual cores the gold standard of semantic processing? the context-dependence of spatial meaning in grounded congruency effects. Cogn Sci 39:1764–1801

Machery E (2009) Doing without concepts. Oxford University Press, Oxford

McRae K, Nedjadrasul D, Pau R, Lo BP-H, King L (2018) Abstract concepts and pictures of real-world situations activate one another. Top Cogn Sci 10:518–532

Meyer K, Damasio A (2009) Convergence and divergence in aneural architecture for recognition andmemory. Trends Neurosci 32:376–382

Nersessian NJ (2002) The cognitive basis of model-based reasoning in science. In: Carruthers P, Stich S, Siegal M (eds) The cognitive basis of science. Cambridge University Press, Cambridge, pp 133–153

Nersessian NJ (1999a) Model-based reasoning in conceptual change. In: Magnani L, Nersessian NJ, Thagard P (eds) Model-based reasoning in scientific discovery. Kluwer Academic/Plenum Publishers, New York, pp 5–22

Nersessian NJ (1999b) Creating scientific concepts. The MIT Press, Cambridge

Pecher D, Boot I, Van Dantzig S, Madden CJ, Huber DE, Zeelenberg R (2011) The sound of enemies and friends in the neighborhood: phonology mediates activation of neighbor semantics. Exp Psychol 58:454–463

Pecher D, Raaijmakers JGW (1999) Automatic priming effects for new associations in lexical decision and perceptual identification. Q J Exp Psychol 52A:593–614

Pecher D, Zeelenberg R, Raaijmakers JGW (1998) Does pizza prime coin? perceptual priming in lexical decision and pronunciation. J Mem Lang 38:401–418

Pecher D, Zeelenberg R, Wagenmakers E-J (2005) Enemies and friends in the neighborhood: orthographic similarity effects in semantic categorization. J Exp Psychol Learn Mem Cogn 31:121–128

Simmons WK, Ramjee V, Beauchamp MS, McRae K, Martin A, Barsalou LW (2007) A common neural substrate for perceiving and knowing about color. Neuropsychologia 45:2802–2810

Van Dantzig S, Raffone A, Hommel B (2011) Acquiring contextualized concepts: a connectionist approach. Cogn Sci 35:1162–1189

Whitney P, McKay T, Kellas G, Emerson WA (1985) Semantic activation of noun concepts in context. J Exp Psychol Learn Mem Cogn 2:126–135

Yee E, Ahmed SZ, Thompson-Schill SL (2012) Colorless green ideas (can) prime furiously. Psychol Sci 23:364–369

Yeh W, Barsalou LW (2006) The situated nature of concepts. Am J Psychol 119:349–384

Zeelenberg R, Pecher D, Shiffrin RM, Raaijmakers JGW (2003) Semantic context effects and primingin word association. Psychon Bull Rev 10:653–660



Graphs in Linguistics: Diagrammatic Features and Data Models

Paolo Petricca(✉)

Department of Modern Languages, Literatures and Cultures, University "G. D'Annunzio" of Chieti-Pescara, Pescara, Italy
paolo.petricca@unich.it

Abstract. This paper examines the use of diagrams in linguistics, in order to analyze their diagrammatic features, as well as their data structures, in relation to the modeling process. It starts from defining the comparison parameters, from several seminal works on diagrams.

The analysis begins from the classical IPA Consonants chart and its relative vowel graph; both are instances of quite orthodox use of visual representations, showing themselves as a mere visual account of relational data. In the syntax section, there is a strong presence of trees with their hierarchical structure.

Semantics offers a far more varied collection of diagrams. Based on many possible foci and on different tasks, there are several important diagrams in use. It will be considered Wordnet, Framenet, Generative Lexicon, DRT and UML-based semantics.

All these examples will lead us to stress how data models play an essential role in determining the modelling potential of graphs and the possible tasks performed by means of them. The object-oriented paradigm and its graphical development will be asserted as the most interesting use of diagrams; it will display an outstanding potential in modelling tasks.

Keywords: Diagrams · Linguistics · Data models · Model-based reasoning

1 A Framework for Diagram Evaluation

1.1 Why Diagrams and Linguistics?

In the last four decades of research, we have shown an increasing interest for diagrams meant as tools for representation and reasoning, with two natural foci: semiotics and logic [12]. Semiotics has investigated the elements of representation in diagrams and how they interact in producing a well-defined group of graphical items capable of modeling almost everything; logic has pointed out the possibility of performing reasoning abilities through diagrams.

In this scenario, linguistics has played an important role, even though it isn't widely acknowledged. In fact, at a first glance, there is an intuitive opposition between diagram and language: usually something that is expressed in language and results unclear to the reader, could benefit of a "translation" into a graphical shape. This opinion stems from the instinctive and naïve opposition of language and drawing, in which diagrams play a middle role.

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 482–499, 2019.
https://doi.org/10.1007/978-3-030-32722-4_27

In the last years, this opposition has been overtaken by the rising of a joint interest of these two fields of enquiry. Of course linguistics has played its part in the study of diagrams as forms of communication, importing its structural modularity (most of all syntax, semantics and pragmatics) and theoretical rules in the study of visual design and diagrams [4, 13]. However, the use of diagrams in linguistics is still not considered central and has received very limited attentions even in these thriving times for diagrams studies [14, 19, 2].

This paper starts from the hypothesis that the use of diagrams in linguistics can prove very useful for several purposes: education, language analysis, theoretical modeling. The aim of this paper is to show how these purposes are connected to several diagrammatic features and to data structures. After the analysis of some typical use of diagrams in linguistics, we will observe how the modeling ability of diagrams depends on data structures and how we decide to manipulate data.

1.2 Limiting the Field

Since diagram studies are very heterogeneous and cover many internal topics, this study appraises as essential an explicit limitation of the used parameters. The first group of choices will be taken into consideration in this part of the paper, while other important features will be discussed later, after being employed in practical analysis.

As the use of diagrams in linguistics is very pervasive and widespread, it is very important to discard some variables, which would not be relevant for our purposes. Therefore, three studies will be used, expressively dedicated to classifying the type of diagrams and their features: [2, 14, 19].

Diagrams are used both in *monolingual* and *multilingual* projects and the latter are aimed at assisting translation processes: translation is one of the most difficult tasks we can perform with language and implies a duplication of all language modules, plus some dynamic interactions. With the specific intention of limiting this complexity, we will treat only the monolingual use of diagrams.

Usually we deal with single modules of language, with concern on phonology and syntax. The situation is less clear with semantics, because whether it is word, phrase or discourse to be represented with diagrams, it is almost impossible to completely divide semantics from syntax and external knowledge. It is precisely because this relationship is naturally active in language, and not just an outcome of our choice, that we can consider this study as concerning *monomodular* diagrams.

Language analysis can interact with both qualitative and quantitative data; as we will see in Part 2, intuitively we can assess the representation of numerical data as a peculiar family of diagrams, directed at organizing and accounting numerical aspects, in an advantageous shape. Therefore, wanting to focus on the modeling skills of diagrams in linguistics, after a brief mention of a numerical use in Part 2, we will consider only *qualitative* uses.

The last, but most renowned limitation, deals with the distinction between *symbolic* and *iconic* representation in diagram elements. Directly inherited from Peirce' Semiotics, this distinction is one of the main concerns in diagram studies and has received many formulations [21]. Iconicity can create a semiotic advantage over a symbolic use of diagrams, but even in the peircean theory [3], we do not have a precise tool of

distinction, capable of assigning every piece of the diagram to a neat iconic or symbolic representation. Then, as we will see, many of the analyzed cases make no use of strongly iconic representation, expressing a clear prevalence of the symbolic one. For these two reasons, we decide to limit the use of this distinction only to the relevant occurrences.

1.3 Devising a Framework

In order to conceive a taxonomy for diagrams and their properties, a common distinction arises: the one between structural and functional aspects of diagrams. While the structural aspects explore the graphic vocabulary actually used, the shapes and pictures, the graphical structures and the mode of representational correspondence, the functional aspects gather information about the interaction with the users and their mental representation and conceptualization, the purpose of the specific representation and the connection with the social context.

In this study, we will focus on the functional aspects of diagrams, especially on the mode of representation and the modeling skills, related to the interaction between user and diagrams. Within a content-related orientation, starting from the data structures of the represented information, we will outline the relationship between the intended tasks, interactions and possible observational advantages [20], in some representative uses of diagrams in phonology, syntax and semantics.

With regard to the mode of representation, following the distinctions suggested by [14, 67–68], we will give attention to the visual features, distinguishing the occurrences in which they function as explicit syntactic or semantic elements, from others in which they are just aesthetics features, enhancing the diagrams comprehension. Following the distinctions in Table 1, we can analyze visual features in strict relation with their tasks and with the position in which they appear.

Table 1. A short outlook of the terminology about visual features.

	Adding visual elements	Visual element placed on the plane	Specifying the format of the visual elements
Purpose of semantics	Embellishments	Notational layout	Notational aesthetics
Purpose of supporting interpretation	Annotations	Layout	Aesthetics

Some separate considerations are made in relation to the opposition between static and dynamic diagrams; most of the diagrams represent static situations and information, but they can contain some dynamic elements (like arrows, or functors) and even interactions between several diagrams or passages from different states of a diagram. Evidently, this possibility is strongly related to the possible tasks and how the users could interact with the graph.

1.4 Data Structures and Data Models

Many strategies for classifying and organizing data types are conceivable and the more we approach to atomic and simple data the fitter would result the mathematical data types; but if we have to deal with a general type of data or aggregate data, the best classification is the one taken from informatics: they are called *Data Structures*. As defined by the Encyclopedia Britannica, data structures are the *≪way in which data are stored for efficient search and retrieval≫* .

Naturally there are many Data Structures, actually in use in ICT, as arrays, queues, records, stacks, linked lists, matrices, graphs et al. Notwithstanding this potentially long list, in linguistics there is a smaller variety of data structures in use and we will encounter most of them. Besides this classification, data are encoded and organized in *data models*, i.e. ways of modeling an information system, aimed at optimizing standardization, interaction between data and the actions performed by the users. Even in this class, a long list of possible models can be found; furthermore, in informatics, the two phrases are sometimes used in an ambiguous, interchangeable way.

In order to try to simplify this classification, and make it reliable and concise for our purpose, it can be considered a valid alternative, the use of the principal conceptual models of database (all definable as data models): hierarchical, relational, and object-oriented.

The *hierarchical* structure is the simplest one and is built on records organized in a tree structure. In this case, the diagram is not just a representation, but also a part itself of the data model considering that diagrams are largely used for hierarchical models, both in informatics and in other fields. The main property of this model is the mono-directional reading, suggested by its typical shape, from the root, at the top, to the leaves, at the bottom. This order can be used, in diagrams, for representing a serial application of a relation (dependency, inclusion, derivation, etc.) and, in some cases, for making the node (or leaf) inheriting the properties of its parental element.

The *relational* structure does not define a single parameter data, but rather a tuple; typically represented as a table, where rows are tuples or a list of values characterizing a single record through each of its parameters (or fields); columns represent a single field, listing all the values for each record. This model allows many operations on records and can be queried in many ways; this versatility constitutes its main advantage.

The *object-oriented* model is based on a set of objects, i.e. an aggregation of data with ontological relevance, possessing a definite identity and certain characteristics. Moreover, objects are considered to interact and communicate with each other, and some of them can incorporate, create or delete other objects. The innovation in this data structure lays in the fact that data are represented within the frame of their object and both the structural and dynamic or interaction data are grouped in this entity. This approach allows a very complex data modeling, enhancing the descriptive power of the whole architecture. This expressive power is counterbalanced by the complexity, both cognitive and computational, of the model.

2 Diagrams in Phonetics

IPA Consonants Chart

Phonetics is probably the most standardized field of linguistics, since most of the phones are classified and physically defined from the International Phonetic Alphabet (IPA). The IPA Charts are very familiar to many scholars and students, especially the pulmonic consonants chart, where places and modes of articulation combine 59 phones made by obstructing the glottis or partially closing the oral cavity and either simultaneously or subsequently exhaling.

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Fig. 1. IPA consonants (pulmonic) chart

This chart contains the vast majority of consonants[1], derived from Latin alphabet, with a very ordinate use of space; shaded areas are *notational aesthetics* representing the sounds considered impossible for our phonatory system. We can see, from Fig. 1, as rows designate the manner of articulation and columns express the place of articulation; as written in the original table caption, phones in double position stand for voiced and voiceless variables.

The iconicity in the chart is small because there are many possible points and manners of articulation, and the latter do not have a topological locus; hence we cannot configure an iconic layout, as we will see in the vowel chart (Fig. 2). It can be produced a partial iconicity, for some of the sounds, but with some necessary *embellishment* and *layout* expedients, as seen in many educational versions of the chart; the same problems would remain in a 3D representation.

These static data are sounds with two or three parameters, with no links or heritance relationships; so a simple chart is naturally representing these relational data, with a common mode of representation. The usual task is to gather and organize those sounds, with particular usefulness for educational and scientific purposes.

[1] There are some extensional charts organizing non-pulmonic and co-articulated consonants.

IPA Vowel Chart

The analogous IPA Vowel Charts [10] presents the same type of data, but in a more diagrammatic shape: here the chart has not its regular form, but is a trapezium, which iconically represents the oral cavity.

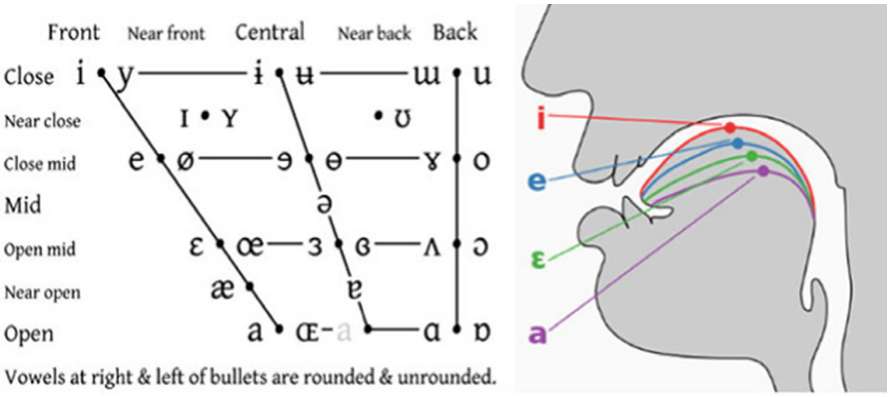


Fig. 2. IPA vowel chart (with iconic mouth correspondence)

As represented in Fig. 2, the trapezium can be read as an oral cavity where the vertical positions are denoting the vowels closeness (top-down), and horizontal positions denote the vowel backness (front at the left of the diagram). This is possible because vowels do not contain differences in voicing, manner, or place; they only differ in the position of the tongue. In this case the *layout* of double sound, means the rounded[2] and the unrounded variables. This chart is complete: no vowels can be pronounced outside this trapezium.

Most languages have a triangular vowel chart, only the 10% use a trapezium, but this difference in layout does not compromise the iconicity of this chart; although this geometric correspondence, the mode of representation remains a common chart, for relational data. The lower number of represented sounds enhances a good observational advantage, and hence the memory supporting and educational purposes of this diagram.

Spectrograms

In phonetics, there is also a frequent use of spectrograms, i.e. the graphical representation of the frequency signals in sounds. They are certainly useful for scientific analysis, but they are not intuitive at all: they need mathematical normalization and a well-prepared eye for their proper, significant use. For these reasons, spectrograms are excluded from this study, like any other possible numerical representation, which exploits a visual shape.

[2] Vowel roundedness is the amount of rounding in the lips during the articulation; in this case the lips form a circular opening while unrounded vowels are pronounced with relaxed lips.

3 Diagrams in Syntax

Syntax deals with the grammatical structure of a sentence and, independently from the possible grammar theories, performs a large use of diagrams, *called Sentence Diagrams*. These diagrams have an important role in giving to speakers an intuitive and organized overview of how phrasal parts compose sentences and under which rules they interact.

Reed and Kellog Diagrams
One of the first, and the historically most prominent use of diagram in syntax, is in the work of Reed and Kellog [6]; they were pioneers in understanding the reliability of diagrams in showing the logical relations in sentence, and in affirming how syntax can be expressed also giving priority to these relations over words order.

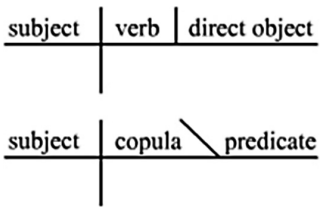


Fig. 3. Reed-Kellog diagram: general rules

Figure 3 shows a general definition of the diagram, where a long horizontal line divides the main constituents of the sentence (in their turn divided by upward vertical lines): subject, verb and object (direct or propositional; in the latter case marked by a backslash). The modifiers (adjectives, articles, adverbs, conjunctions and prepositions) are positioned downward the baselines, all connected with backslashes, possibly nested if they contain a new phrase (nominal or verbal, as shown in Fig. 4). It is thus evident how the *notational layout* plays the most important part in this graph.

The purpose of these diagrams is to illustrate how the linear logical structure of phrase can be preserved despite the many possible expansions and to stress the prominent position of verbs and nouns putting them at the top of the diagram. In this system, there is no relevance for iconicity because the relationship between the Parts of Speech has not a visual nature. The main limit of this diagram is that it is not capable of showing how a single part acts over the phrase, because it divides the structural role of

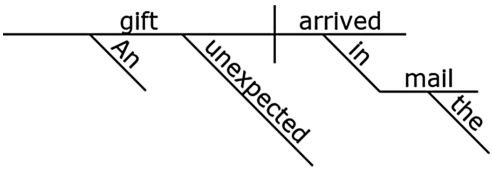


Fig. 4. Reed-Kellog diagram: an application

every single word: e.g. in Fig. 4 the article 'An' is a modifier not just of the subject 'gift', but of the whole noun phrase 'unexpected gift'. Moreover, the Reed and Kellogg's diagram has influenced both constituency and dependency diagrams and has produced even some contemporary hybrid solutions for tree parsing.

Constituency and Dependency Diagrams

Many other diagrams are used in syntax, due to the fact that during the last seventy years several grammar theories have been produced and debated, each rather different from the others. Functionalist, Categorial, Dependency and Generative grammars are respectively intended to describe and rule various grammatical features and organizations.

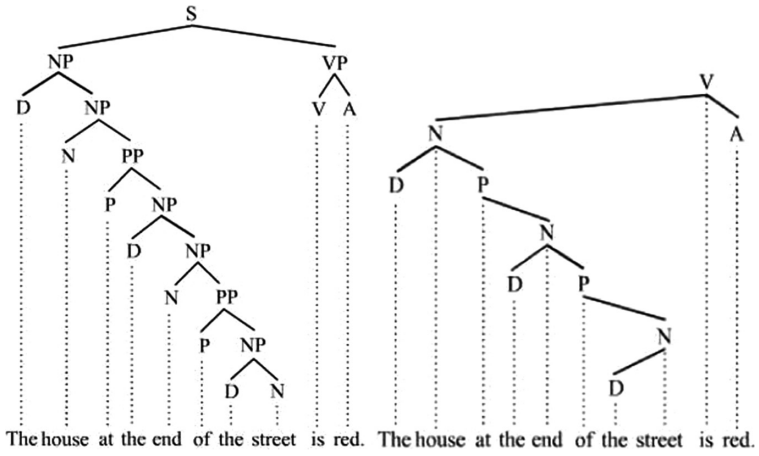


Fig. 5. A constituency grammar (left) and a dependency grammar (right)

This abundance of grammar theories and of possible diagrams makes it impossible to have a full cover in this study, as it will happen for semantics too. For this reason, it is convenient to choose a representative example of typical tree diagrams able to fulfill our framework requests: Constituency and Dependency diagrams are adequate for this aim.

Constituency grammars derived from the works of Bloomfield and Chomsky; their use was strengthened by the formal use of Chomsky's Context-free and Context-Sensitive grammars. The constituency diagrams are typical of all the versions of Generative Grammar even if they present some differences. As in Fig. 5 the root is always identified with the whole sentence (S) and stems through its basic constituents (NPs, VP, and DP), in turn stemming (with possible nesting) to their final constituents (single words and their part of speech).

On the other hand, dependency relation, derived from the work of Tesnière, focuses on the various dependence relations in the sentence that stems directly from the verb (root node): the verb is taken to be the structural center of clause structure. All other syntactic units are, directly or indirectly, connected to the verb, and so dependent from it; there are several kinds of dependencies that can be tagged with traditional grammar phrases initials or with other syntactic denominations.

It is important to report how in the first architecture there are, usually, many one-to-many relations, while in the second one-to-one relations are more frequent: this depends on the modeling perspectives, which are represented in diagrams. Moreover, the dependency tree is an ordered tree, where the order of words remains unaltered, and so puts its focus on the linearity of grammar.

Independently from these theoretical diversities, both the diagrams represent hierarchical data models, mostly by means of *notational layout* and *embellishments*. Phrase structures are expressed by single constituency or dependency rules, which have their propositional expression; hence, diagrams are intended to gather all these applications of rules, in order to have a complete overview of the analysis and of all its single steps. Data can be interpreted both as static or dynamic, and the layout allows an important observational advantage, enhancing the ability of the user of state consistent syntactic relationships, at a glance.

4 Diagrams in Semantics

Semantics is the most complex and articulated field of linguistics; many diagrams in several sectors of this topic can be detected: cognitive, formal, conceptual, and lexical. All these addresses deliberately choose to focus on one or few aspects of language semantic, thus leaving out the remaining ones; for this reason it is necessary a selection of diagrams that exemplifies the important uses in semantics both from a data models and a general modeling purpose point of view.

4.1 Wordnet and Framenet

Wordnet [5] is, probably, the most recognized corpus of lexical semantics and it is based upon the concept of *synset*, i.e. a unity of meaning designing one or more words, considered synonyms. Every synset has a network of semantic relations with other synsets: hyperonymy/hyponymy, holonymy/meronymy, coordination (for nouns); troponymy, hypernymy and entailment (for verbs); antonymy (for adjectives) and other cross-PoS relations.

The result is a database, with each record (word) containing just its verbal definition, and linked to other words, by the above-mentioned relations. In the SQL DB, records can have different numbered senses and several attributes (morphemes, lexical domain, verbal frames, adjectives position, etc.), which basically make this software a relational DB. However, there are also several link types, expressing the semantic relations; since these relations are the main feature expressed by this linguistic tool, that otherwise would just be a dictionary, the only diagram in use for Wordnet is a graph depicting those relations, one at a time, for the sake of visual clarity.

These relations give a combinatorial account of the semantic field and can be expressed by means of an Entity-Relation Graph, which typically represents the only hyperonym/hyponym relation, with the starting synset represented as the root node.

The central information of lexical semantics is here expressed through a diagram of a single semantic relation, which has direct and indirect linking, so it cannot be considered as a simple tree diagram, but as a graph, in the mathematical sense. Here is

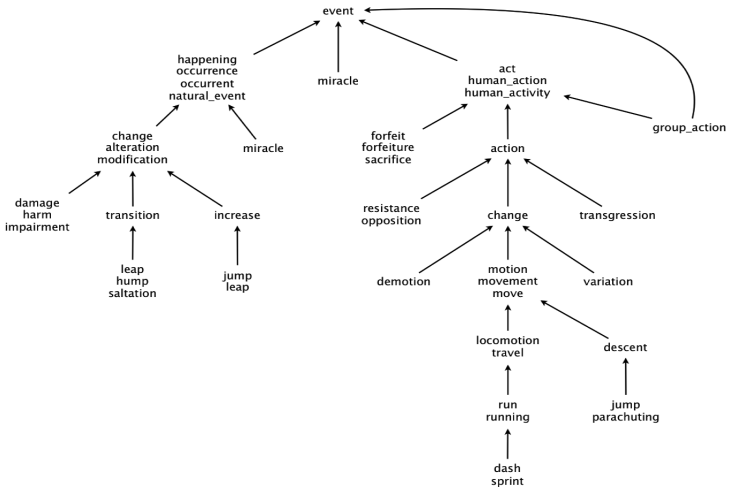


Fig. 6. The Wordnet relational graph for the word *event*.

evident the fact that the high connectivity among synsets allows an easy exploration of semantic relations, making this tool particularly fit for many semantic tasks. Hence these diagrams are valuable for the process that we can call *human information retrieval*; users can easily appraise all the related words and this constitutes a clear observational advantage. Furthermore, it gives an overview (actually with little ambiguity) of the relations between different synsets of the same word, e.g. in Fig. 6. *miracle* and *group_action*.

FrameNet [6, 16] has a different aim: describing all the elements of a semantic frame and their required semantic type; the semantic frame is a sort of automatic semantic structure that the speaker evokes for the instantiated use of a word. This frame is defined by elements, belonging to a semantic type, divided between core and non-core elements; in addition, there are frame to frame relations (Inherits from; Is Inherited by; Perspective on; Is Perspectivized in; Uses; Is Used by; Subframe of; Has Subframe (s); Precedes; Is Preceded by; Is Inchoative of; Is Causative of; See also) and Lexical units (other single word or locutions with identical framing process [synonyms]).

Even if the data model of Framenet is again a relational DB, the performed modeling scheme is considerably different from the one obtained through Wordnet. Here the records are not single words or certain PoS, but cognitive frames containing several different elements, each with its role, relations and specific lexical applications. The model is far more complex than the one in Wordnet, with its type theory, constraints, and dynamic interactions between frames. This complexity does not conform to a single diagram, which would lose all its intuitive features; but the researcher team has found a valid solution.

Frame Grapher is a querying software, devised for generating diagrams from any selected frame; by choosing how many relations, generations and peripheral children have to be represented, the user can configure his desired complexity. The resulting diagram is a graph with many embellishments, necessary for expressing this complex

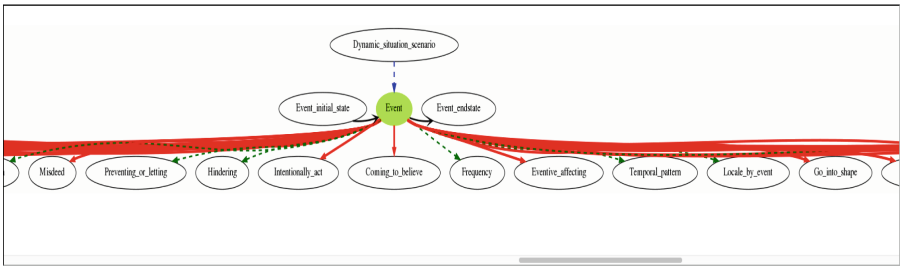


Fig. 7. The Framenet relational graph (Frame Grapher) for the word *event*.

visual grammar. As we can see in Fig. 7, the frame purview, using *notational aesthetic* differences, is composed of many nodes (this screenshot is just a quarter of the entire graph) in different shapes[3] and colors, with some links designed by arrow-shaped arcs, solid and dashed. Data are dynamic and allow a string observational advantage, with the only limitation pertaining the huge dimension that diagram can reach and the consequent low readability.

Here the important aim is to define some basic cognitive operations made by speakers, in understanding semantics and reasoning on it; the graph can give a complete but complex overview to the frames relations, allowing a precise graphical expression of this complex model-based semantics and allowing the users to see at a glance, some key relations in a much more detailed way than in Wordnet.

We can assert then that in these two diagrams, relations are represented with common geometrical shapes, showing a high grade of symbolic use of diagrams; both diagrams are very informative and useful for human tasks that would result much more difficult if executed on propositional or coded relational data. The two theoretical ways of modeling semantics are specular to the respective diagram but, as it will become clearer in the final part of this work, these two diagrams have secondary aims for the two theories and are not devised for a complete representation of the theory nor for a direct modeling purpose.

4.2 Generative Lexicon and Discourse Representation Theory

These two semantic theories, very successful in formal semantics, do not present a use of diagrams directly useful for our purposes, but have an interesting role in exemplifying the relationship between data models and diagrams, in semantics.

Generative Lexicon [15] is a semantic theory focused on the distributed nature of compositionality in natural language, spreading the meaning of utterances throughout all the components. It intends to explain the generative capacity of speakers from a semantic point of view; this aim is reached thanks to a knowledge representation framework that offers a rich and expressive vocabulary for lexical information.

[3] Even if not present in Fig. 7, in the peripheral parts of diagrams there can be square nodes, representing groups of children or sibling frames.

Lexical and general knowledge are both covered by this theory, which tries to model general knowledge about words in four possible articulations: a. *lexical typing structure*: giving an explicit type for a word positioned within a type system; b. *argument structure*: specifying the number and nature of the arguments to a predicate; c. *event structure*: defining the event type of the expression and any sub-eventual structure it may have with sub-events; d. *qualia structure*: a structural differentiation of the predicative force (formal, constitutive, telic and agentive) for a lexical item.

As sketched in Fig. 8, each structure is expressed by formal specification, composing a linear notation conceived to facilitate a formal treatment of lexical knowledge. And hence, the aim of this diagram is to gather all the structures and have a complete overview of the precise interactions between them, for a correct modeling of these words semantics.

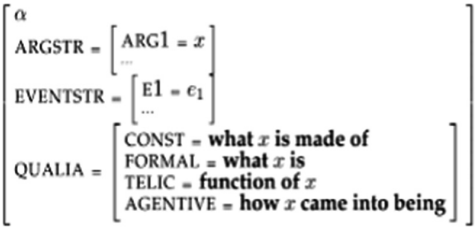


Fig. 8. A general diagram for a GL lexeme

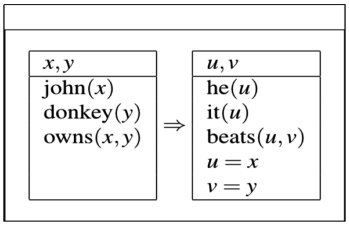


Fig. 9. A DRT diagram

Therefore, we could say that, in this theory, there is a strong modeling attitude, which has produced many important outcomes in recent years, but since the main notation for this theory is formal, the diagram does not produce a real *observational advantage* [20] but a clearly organized *layout* of structures.

With DRT [11] we go briefly out of general lexical semantics, toward formal semantics. This is a high-value formal theory for representing the speakers' mental representation of discourse, particularly effective for the quantifiers' scope problems and anaphoric references. Here too the main notations are formal, deriving from first-order logic, and the diagrams are not isomorphic expressions of this formal system (as Euler diagrams, or Pierce's Existential Graphs), but a boxed notation (Fig. 9) producing an iconic representation of the domain of quantification and predication, able to avoid the possible ambiguities. This leads to a fringe diagrammatic nature, where the static layout features have just the virtue of ordering.

5 Modeling Semantics with UML

This section will take into consideration the Unified Eventity Representation (UER) [17, 18], devised by A. Schalley, as a de-compositional semantics theory. This framework is based on UML [7], a universal modelling language born in 1997, in the engineering sector, with the purpose of graphically modelling any software, before the programmers effectively start to code. It is a set of best practices, very useful for many complex

modelling, which uses its structural modularity, through several diagrams techniques (14 diagrams syntaxes: 7 for structures, 7 for operations), as a clear overview of the division of work and the control of step-by-step creation of a software. It is well known among the Object-Oriented (O-O) community, used especially for the communications with clients and managers, but it is not effectively used by the largest part of O-O coders, mostly because they are accustomed to the formal notations of their domain specific language (in a very similar way to the use Peirce's EG related to First-Order Logic).

The UER has basically the same syntax and the same semantics of UML, but does not use all the 14 modules and most of all it is centered upon the concept of *eventity,* i.e. a conceptualization or scheme of what is happening in a semantic frame, usually relating to its main verb. The orientation of this model is, of course, cognitive: we understand meanings to be conceivable as concepts; this process of conceptualization is strongly related to *non-linguistic knowledge*.

Here it is useful to specify that the previously cited Generative Lexicon theory is an object-based theory itself, bearing nearly the same data structure. Both theories use an object-based architecture, but there are substantial differences: the O-O paradigm uses also *inheritance* (which allows the creation and use of a complete ontology and taxonomy, using the structural properties of classes, attributes, constraints, etc.) and *polymorphism* (i.e. the ability to process objects differently, depending on their data type or class, so that different types can be accessed through the same interface).

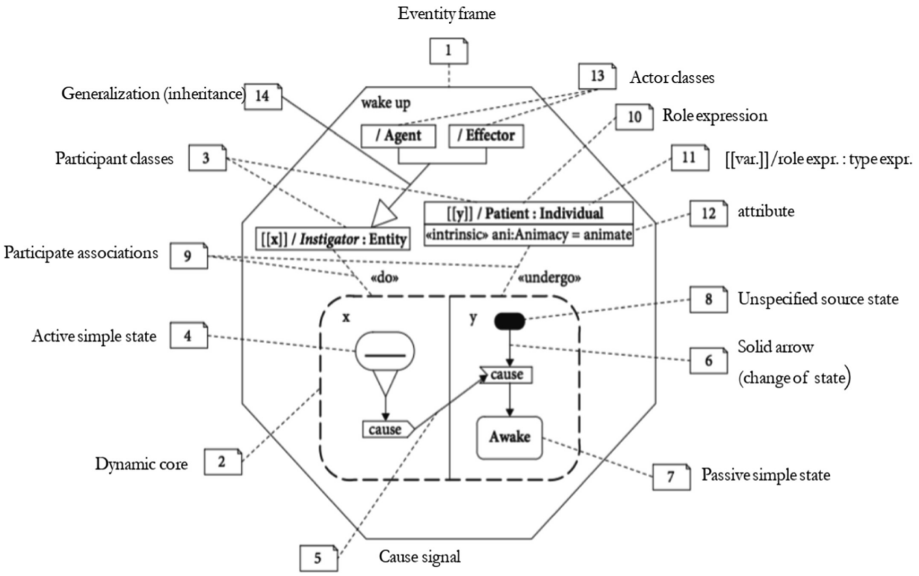


Fig. 10. Eventity diagram for *wake up*

This diagrammatic theory is complex, modular and articulated. Here it will be considered only a classic introductive example, further analyzing the key features of this tool. UER's main class of diagrams is the Eventity Diagram (Fig. 10).

The frame is divided into two sectors: Participants x and y of the eventity are represented by rectangles (3); the dynamic core (2) is defined by a dashed-line square with round corners. The actors are represented in their class diagrams (3) defined by roles, type, attributes and their eventual taxonomic relations (14). The role-expression (10) names the semantic role, while the type-expression (11), lists the ontological category of the participant, restricting its type. For the undergoer of the example, a displayed selectional restriction can be seen: it is specified as an animate entity by way of an attribute (12). In the actor x's participant class specification, the role-expression Instigator is italicized, indicating that it is an abstract characterization that cannot be directly instantiated. Actors are instead either *Agents* or *Effectors* (13), as expressed by the generalization or inheritance relationship (14). They inherit all the properties of Instigator with the additional information of whether the actor acts volitionally (*agent*) or non-volitionally (*effector*).

In the dynamic part, active simple states (4) are containers with straight top and bottom and convex arcs on the sides; the underscore in its body illustrates that it is not conceptualized what x is doing, so the action is unspecified. Causation is expressed in the UER by the cause-signal (5), consisting of the two pentagons (the convex one in x's *swim-lane* indicates the sending of the signal, the concave one in y's *swim-lane* the receipt) and the connecting arrow. The y's change of state is represented by a transition shown in y's *swim-lane*; this transition comprises a solid connecting arrow (6), a source and target state. The target state in the *wake up* eventity is a passive simple state (7), depicted by a rectangle with rounded corners, which is specified as *Awake*. The source state is unknown and so displayed as an unspecified source state (8). Dashed connectors (9) link the participant classes and the dynamic core.

This brief overview of the composition of diagram let us appraise how powerful and complex this modeling language can be; in this diagram (rather simple for the expressivity of UER) is expressed the knowledge representation of the speaker's mental frame of *wake up*, including the object-related class and type theory, taxonomical attributes of the components and a modular diagram for the dynamic part of semantics with its own considerable descriptive power. The abundant structural syntax of UER, comprises also: a constraint module based in OCL language, the possibility of using enumeration as the extensional definition of a class, a list of structural properties, the use of stereotypes (creating new classes from a single instance and its features), aggregations, associations and a rich dynamic structure[4].

This remarkable variety of modeling instruments is devised for obtaining a *variable grade of granularity* and an important *cognitive adequacy*. When speakers try to recall all the concepts necessary for a sound semantic performance, they do not recall entire taxonomies or ontologies, but the necessary amount of data; this process is not performed in a strict formal way nor following a universal structure. The human reasoning about utterances and discourses follows our network-shaped mind, picking up useful concepts (usually starting from a single object evoking a frame) following intuition, stereotypes, abduction and, of course, our own pertinent knowledge [8, 9]. UER, more

[4] Schalley [18] introduces also some pre-created eventity models able to describe classic interactive situations using a graphical notation for describe states in time, actions, changes, etc.

than every other theory considered here, tries to be accurate in following this human process; the result is a framework characterized by high cognitive fidelity and an intuitive graphical shape. Notwithstanding these efficiencies, UER has its feet of clay in its technical profile, because it inherits the complexity of UML (even if strongly reduced) and uses a wide graphical alphabet.

Key Technical and Diagrammatic Features

The most important feature of O-O paradigm, and of UER, is *encapsulation*: it is the possibility of jointly representing, in a unique graph, characteristics (attributes), relation with other objects, behaviors (actions, states, change of states), interactions (information passing and decision making). The information contained in the object is hidden to the outside sectors of the code and can be manipulated only by processing this specific object. This is the main similarity with the human cognitive process: an object is an instance that exists at *execution time*[5] and follows its class definition, while the corresponding class is a salient concept, to which belong similar objects, with shared characteristics and dynamics. So classes follow a *type theory* and general rules, already established and complete, while objects are both following their class constraints and being defined in execution time, while minds or computers are running it.

This happens also in linguistic semantics and cognition: an object, i.e. a discourse referent with all its related knowledge (e.g. an entity), while it is a member of a context-independent ontological category, can be used in a singular way or being associated, just for a single semantic evaluation, to a peculiar context-sensitive behavior or rule. Our cognitive system [22, 23] is centered on categorized entities (single concepts from different levels of an ontology) and on how we primarily define them, how they relate to each other and interact, and what are their roles, rules and behavior in dynamic processes. This correspondence shows how the O-O data model is the most expressive and adequate model for representing linguistic semantics: UER can be considered as the apex of cognitive modeling in this branch of linguistics.

UER has a *multi-layer semantics*; ontological definitions of static data are achieved through a multi-layered architecture, consisting of four distinct and ordered modeling layers: *instance layer* (user objects, e.g. *my car* belongs to this stratum), *model layer* (where *my car* is an instance of the *Car* class), *metamodel layer* (where the *Car* class is an instance of the Class-of-classes), *meta-metamodel layer* (where a generic Class-of-classes is defined through Meta-Class attributes). This feature allows a highly complex type theory and an impressive descriptive power, but at the same time brings a severe computational complexity, for class diagrams [1]. Nonetheless, since UER is theoretically far from the formal semantics aims, is not significantly affected by this problem; still, from a diagrammatic perspective, this complexity can be a solid obstacle for the users.

The UER diagrams are characterized by a strong use of topology (for encapsulation, hierarchical relation and dynamics) and more specifically of *notational layout*, besides a rich lexicon of *embellishments*. The overall mode of representation is symbolic with a strong effort in using simple, conventional and universal signs; there is as well a frequent use of textual elements. The innate representative nature of this theory brings

[5] It is not previously coded in a higher-level structure.

with it the highest grade of description, hence all the information of a single diagram are fully contained in it. The possible inferences are related to class-diagram inheritance, encapsulation and the modularity of diagrams: some data can be partially represented in a graph and completely in another, allowing possible inter-graphs inferences. This reasoning is formally complex, a full formal study about the subject is not yet available and the inference engines coded concern the sole class diagram. This mix of static and dynamic modeling is very expressive in its diagrammatic shape and gives plenty of observational advantages, for every module.

6 Data Models and Diagrammatic Tasks

After completing the detailed survey of these uses of diagrams in linguistics, we can focus on the relations between diagrammatic features and data models. This allows us to state several specific considerations with evaluative and proposing range.

In Phonetics, diagrams are used as means of account for the phonemes organization. The consonants chart shows a rational layout for organizing the phonetic production and gives an effective overview, useful for learning and remembering all the consonants. However, this is possible mainly thanks to the restrict number of consonants and the natural relational description of these entities. For the vowels we have seen a rather similar situation with the additional possibility of iconically depicting the place of articulation. Both these uses of diagrams do not involve a consistent modeling task, since the represented data are already structured in their data model, in their non-diagrammatic versions.

Syntax diagrams display a similar situation as regards to the preceding contribution of data model: syntactic structures are usually defined in a tree-hierarchical shape. The use of diagrams shows with clearer evidence this structure and allows the choice of a theoretical focus on the kind of relationships relevant for each and every study. Only in this latter option we can ascertain a modeling act; this task comes with a null iconic value. As a matter of fact, there is a similar scientific value for formal and linear accounts of syntax in respect of diagrammatic ones, the latter being valued as more productive for educational purposes.

In semantics we have encountered a much more varied and interesting situation. In corpora semantics data are organized in some relational databases, and queried with auxiliary softwares; diagrams are not native for these theories and are later devised for offering an easy human consultation and the analysis of relevant relations. The modeling activity is completed in the pre-diagrammatic stage, where data are collected within the function-argument structure. Framenet, with its more complex model, shows a sensitive cognitive effectiveness in semantic modeling and its resulting graphs have some interesting features, but the use of symbolic notation, with the different colored arrows, requires for the user a certain familiarity with the possible frame relations. Beyond this potential complication, the graph shows a good overview of the frame related concepts, necessary for the complete semantic interpretation of the utterance.

In DRT, the graphical version has just an organizing purpose, since the modeling process is expressed in logical form, as in most of the formal semantics theories; here diagrams have a very limited function. Generative Lexicon has a similar use of

diagrams focused on grouping and organizing information; hence, even if the theory has got a relevant modeling value for linguistic semantics, we cannot say the same of its graphical shape.

The UER framework shows the most important potential for the use of diagrams in linguistics. Based on a native diagram modeling language, it expresses the full modeling power of diagrams. In UER, cognitive semantics finds its most expressive model, capable of giving a formal, and yet graphical, account of how ideally speakers use their knowledge, in order to build a complex semantic representation of verbs. The modeling potential of this approach is impressive, as testified by the reputation of UML; such extended modeling language can be shaped for many modeling aims, always considering the inevitable tradeoff between expressivity and complexity. The cognitive adequacy of this language confirms the strategic role of vision and spatial layout in the conceptualization and organization of data.

It is important here to remark the fact that, even if we cannot have a reliable and complete set of semantic primitives, and diagrams do not even seem to be an effective stand-alone tool for the representation and modeling of linguistic semantics, the use of diagrams for cognitive semantics modeling still appears as the highest potential diagrammatic task in linguistics.

7 Conclusions

The present study shows how the modelling potential of diagrams in linguistics can be considered as directly proportional to the complexity of data model represented. When data can be organized in taxonomies or hierarchies, diagrams lose some modelling abilities, while object-based data offer the widest scope both in representing and modelling with diagrams.

The UML language showed consistent potentialities in modelling problems and knowledge, with high observational advantage. The exploration of these potentialities is still incomplete and can be extended to the non-verbal semantics and the inferential possibilities of UML structural diagrams, other than Class Diagrams.

References

1. Berardi D, Calvanese D, De Giacomo G (2005) Reasoning on UML class diagrams. Artif Intell 168:70–118
2. Blackwell A, Engelhardt Y (2002) A meta-taxonomy for diagram research. In: Anderson M, Meyer B, Olivier P (eds) Diagrammatic representation and reasoning. Springer, London, pp 47–64
3. Eisele C (ed) (1976) The new elements of Mathematics by Charles S. Peirce. Mouton, The Hague
4. Engelhardt Y (2002) The language of graphics: a framework for the analysis of syntax and meaning in maps, charts and diagrams. Ph.D. thesis, Universiteit van Amsterdam, Institute for Logic, Language and Computation
5. Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT Press, Cambridge https://wordnet.princeton.edu/

6. Fillmore C et al (2003) Background to FrameNet. Int J Lexicography 16(3):235–250
7. Fowler M, Scott K (1997) UML distilled - applying the standard object modeling language. Addison Wesley, Boston
8. Geeraerts D (2010) Theories of lexical semantics. OUP, Oxford
9. Helbig H (ed) (2006) Knowledge representation and the semantics of natural language. Springer, Berlin
10. IPA Homepage. https://www.internationalphoneticassociation.org/content/full-ipa-chart. Accessed 28 Feb 2019
11. Kamp H, Reyle U (1993) From discourse to logic. introduction to model theoretic semantics of natural language, formal logic and discourse representation theory. Studies in linguistics and philosophy, vol 43. Springer, Netherlands
12. Kramer S, Ljungberg C (eds) (2016) Thinking with diagrams. The semiotic basis of human cognition. De Gruyter Mouton, Berlin
13. Kress GR, van Leeuwen T (1996) Reading images: the grammar of visual design. Routledge, Abingdon
14. Purchase HC (2014) Twelve years of diagrams research. J. Visual Lang. Comput. 25(2):57–75
15. Pustejosvky J (1995) The generative lexicon. MIT Press, Boston
16. Ruppenhofer J, Ellsworth M et al (2016) FrameNet II: extended theory and practice. https://framenet.icsi.berkeley.edu/fndrupal/
17. Schalley AC (2004) Representing verbal semantics with diagrams. an adaptation of the uml for lexical semantics. In: Proceeding COLING 2004 proceedings of the 20th international conference on computational linguistics article no 785. Association for Computational Linguistics, Stroudsburg
18. Schalley AC (2004) Cognitive modelling and verbal semantics. De Gruyter Mouton, Berlin
19. Smessaert H, Demey, L (2018) Towards a typology of diagrams in linguistics. In: Chapman P, Stapleton G, Moktefi A, Perez-Kriz S, Bellucci F (eds) Diagrammatic representation and inference. diagrams 2018. Lecture Notes in Computer Science, vol 10871. Springer, Cham, pp 236–244
20. Stapleton G, Jamnik M, Shimojima A (2017) What makes an effective representation of information: a formal account of observational advantages. J Logic Lang Inform 26 (143):143–177. https://doi.org/10.1007/s10849-017-9250-6
21. Stjernfelt F (2011) On operational and optimal iconicity in Peirce's diagrammatology. Semiotica 186(1/4):395–419. https://doi.org/10.1515/semi.2011.061
22. Talmy L (2000) Towards a cognitive semantics. vol. 1 concept structuring. MIT Press, Cambridge
23. Talmy L (2000) Towards a cognitive semantics. vol. 2 typology and process in concept structuring. MIT Press, Cambridge

Author Index

A
Arfini, Selene, 41

B
Bagassi, Maria, 120
Beni, Majid D., 153
Benzmüller, Christoph, 441

C
Caiani, Silvano Zipoli, 427
Caravona, Laura, 120
Casacuberta, David, 318
Caterina, Gianluca, 256
Cortés-García, David, 138
Cucchiarini, Veronica, 120

D
Dellantonio, Sara, 407
Drago, Antonino, 298

F
Ferretti, Gabriele, 173, 427
Fuenmayor, David, 441

G
Gangle, Rocco, 256
Gaytán, David, 372
Gelfert, Axel, 3
Giovagnoli, Raffaela, 55
Giunti, Marco, 20
Goodwin, William, 245

H
Hermkes, Rico, 274

I
Ippoliti, Emiliano, 393

L
Lagnado, David, 103
Li, Ping, 470
Lopez-Orellana, Rodrigo, 138

M
Macchi, Laura, 120
Mach, Hanna, 274
Magnani, Lorenzo, 217

P
Pastore, Luigi, 407
Pereira, Luís Moniz, 69
Petricca, Paolo, 482
Pinna, Simone, 20

R
Rivadulla, Andrés, 287

S
Sans, Alger, 318
Santos, Francisco C., 69

© Springer Nature Switzerland AG 2019
Á. Nepomuceno-Fernández et al. (Eds.): MBR 2018, SAPERE 49, pp. 501–502, 2019.
https://doi.org/10.1007/978-3-030-32722-4

T
Tohmé, Fernando, 256

V
Veit, Walter, 83
Viola, Marco, 173
Vorms, Marion, 103

W
West, Donna E., 193
Woods, John, 337

X
Xiang, Bideng, 470