

Lorenzo Magnani  
Ping Li (Eds.)

STUDIES  
IN APPLIED  
PHILOSOPHY,  
EPISTEMOLOGY  
AND  
RATIONAL  
ETHICS

**SAPERE**

# Philosophy and Cognitive Science

Western & Eastern Studies

 Springer

## Editor-in-Chief

Prof. Dr. Lorenzo Magnani  
Department of Arts and Humanities  
Philosophy Section  
University of Pavia  
Piazza Botta 6  
27100 Pavia  
Italy  
E-mail: [lmagnani@unipv.it](mailto:lmagnani@unipv.it)

## Editorial Board

Prof. Atocha Aliseda  
Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México (UNAM),  
Ciudad Universitaria, Coyoacan, 04510, Mexico, D.F.  
E-mail: [atocha@filosoficas.unam.mx](mailto:atocha@filosoficas.unam.mx)

Prof. Giuseppe Longo  
Laboratoire et Département d'Informatique, CREA, CNRS and Ecole Polytechnique, LIENS,  
45, Rue D'Ulm, 75005 Paris, France  
E-mail: [Giuseppe.Longo@ens.fr](mailto:Giuseppe.Longo@ens.fr)

Prof. Chris Sinha  
Centre for Languages and Literature, P.O. Box 201, 221 00 Lund, Sweden  
E-mail: [chris.sinha@ling.lu.se](mailto:chris.sinha@ling.lu.se)

Prof. Paul Thagard  
Department of Philosophy, Faculty of Arts, Waterloo University, Waterloo, Ontario,  
Canada N2L 3G1  
E-mail: [pthagard@uwaterloo.ca](mailto:pthagard@uwaterloo.ca)

Prof. John Woods  
Department of Philosophy, University of British Columbia, 1866 Main Mall BUCH E370,  
Vancouver, BC Canada V6T 1Z1  
E-mail: [john.woods@ubc.ca](mailto:john.woods@ubc.ca)

Lorenzo Magnani and Ping Li (Eds.)

---

# Philosophy and Cognitive Science

Western & Eastern Studies

 Springer

*Editors*

Prof. Dr. Lorenzo Magnani  
Department of Arts and Humanities  
Philosophy Section  
University of Pavia  
Piazza Botta 6  
27100 Pavia  
Italy  
E-mail: [lmagnani@unipv.it](mailto:lmagnani@unipv.it)

Prof. Ping Li  
Department of Philosophy  
Sun Yat-sen University  
No.135, XinGang Xi Lu  
Guangzhou 510275  
P.R. China  
E-mail: [hsslip@mail.sysu.edu.cn](mailto:hsslip@mail.sysu.edu.cn)

ISSN 2192-6255

e-ISSN 2192-6263

ISBN 978-3-642-29927-8

e-ISBN 978-3-642-29928-5

DOI 10.1007/978-3-642-29928-5

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012936807

© Springer-Verlag Berlin Heidelberg 2012

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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

Today, the relationships between Asia and the Western world make the headlines only when they concern economic deals, folk-ideological confrontations, or divergent ideas on how to solve international crises. The cultural and, more specifically, academical links are frequently disregarded. This book aims at being an argument against such systematic lack of interest for the results of collaborations between Western and Eastern intellectuals and academics: what emerges from the juxtaposition of papers of different geo-cultural origins – but dealing with the same issues – is sometimes a novel approach, which takes advantage of the multifaceted sensibilities inherited by the scholarly legacies who contributed to the debate. This volume is a collection of selected papers that were presented at the international conference *Philosophy and Cognitive Science* (PCS2011), held at Sun Yat-sen University, Guangzhou, P. R. China in May 2011.

Previous volumes prepared the basis for the realization of PCS2011, as a conference explicitly devoted to the conjunction of western and eastern studies. Those volumes also originated from international joint research projects, which succeeded in establishing a first relationship between the two worlds in the area of philosophy and cognitive science. *Model-Based Reasoning in Scientific Discovery*, edited by L. Magnani, N.J. Nersessian, and P. Thagard (Kluwer Academic/Plenum Publishers, New York, 1999), based on the papers presented at the first “model-based reasoning” international conference, held at the University of Pavia, Pavia, Italy in December 1998, has been translated in Chinese, China Science and Technology Press, Beijing, 2000. *Abduction, Reason, and Science* by L. Magnani was translated by Dachao Li and Yuan Ren and published by Guangdong People’s Publishing House, Guangzhou, in 2006. Other volumes, *Science, Cognition, and Consciousness*, edited by P. Li *et al.* (JiangXi People’s Press, Nanchang, China, 2004, published in Chinese and English), *Philosophical Investigations from a Perspective of Cognition*, edited by L. Magnani and P. Li (Guangdong People’s Publishing House, Guangzhou, China, 2006, published in Chinese), *Model-Based Reasoning in Science, Technology, and Medicine*, edited by L. Magnani and P. Li (Springer, Berlin/New York, 2007), derived from the following previous conferences: “Model-Based Reasoning in Science and Medicine” (MBR06\_CHINA, held at Sun Yat-sen

University, Guangzhou, China, July 2006), and the first “Philosophy and Cognitive Science” international conference (PCS2004, held at Sun Yat-sen University, Guangzhou, China, June 2004).

The presentations given at the Guangzhou conference addressed various recent topics at the crossroad of philosophy and cognitive science, especially taking advantage of both western and eastern research. The selected papers contained in the proceedings mainly focus on the following areas: abductive cognition, visualization in science, the cognitive structure of scientific theories, the nature and functions of models, scientific representation, mathematical representation in science, model-based reasoning, analogical reasoning, moral cognition, cognitive niches and evolution. Three symposia characterized the workshop – Symposium on Scientific Representation: Theories and Models (Chairs: Zhilin Zhang and Zhenming Zhai); Symposium on Abductive Cognition (Chair: Lorenzo Magnani); and Symposium on Culture and Cognition (Chairs: Remo Job and Jing Zhu). The various contributions of the book are written by interdisciplinary researchers who are active in the area of philosophy and/or cognitive science.

The editors wish to express their appreciation to the members of the Scientific Committee for their suggestions and assistance:

- Akinori ABE, Innovative Communication Laboratory, NTT Communication Science Laboratories, Kyoto, JAPAN - Liliana ALBERTAZZI, Faculty of Cognitive Science & Centre for Mind and Brain, University of Trento, Rovereto, ITALY
- Emanuele BARDONE, Department of Philosophy, University of Pavia, Pavia, ITALY - Xiang CHEN, Department of Philosophy, California Lutheran University, Thousand Oaks, CA, USA - Xiaoping CHEN, Institute of Philosophy, South China Normal University, Guangzhou, CHINA - Ronald N. GIERE, Center for Philosophy of Science, University of Minnesota, USA - Michael E. GORMAN, Division of Social and Economic Sciences, National Science Foundation, Arlington, VA, USA
- Remo JOB, Faculty of Cognitive Science, University of Trento, Rovereto, ITALY
- Ping LI, Department of Philosophy, Sun Yat-sen University, Guangzhou, CHINA
- Xingmin LI, Graduate School, Chinese Academy of Sciences, Beijing, CHINA - Lorenzo MAGNANI, Department of Philosophy, University of Pavia, Pavia, ITALY
- Woosuk PARK, Humanities & Social Sciences, KAIST, Guseong-dong, Yuseong-gu Daejeon, SOUTH KOREA - Claudio PIZZI, Department of Philosophy and Social Sciences, University of Siena, Siena, ITALY - Ryan D. TWENEY, Department of Psychology, Bowling Green State University, Bowling Green, OH, USA
- Yidong WEI, Center for Philosophy of Science & Technology, Shanxi University, Taiyuan, CHINA - Guolin WU, Center for Philosophy of Science & Technology, South China University of Technology, Guangzhou, CHINA - Zhenming ZHAI, Department of Philosophy, Sun Yat-sen University, Guangzhou, CHINA - Huaxia ZHANG, Department of Philosophy, Sun Yat-sen University, Guangzhou, CHINA - Zhilin ZHANG, School of Philosophy, Fudan University, Shanghai, CHINA - Jing ZHU, Department of Philosophy, Sun Yat-sen University, Guangzhou, CHINA.

Special thanks also go to Emanuele Bardone and Tommaso Bertolotti for their contribution in the preparation of this volume and to Lizhen Sun, Fang Yu, Lingyun Yang, Hui Li, and Shenghua Wang for their work in the local organizing committee of the Guangzhou conference. The conference PCS2011, and thus indirectly this book, was made possible through the generous financial support of Sun Yat-sen University, ZhanJiang Chemical Industrial Incorporated Corporation, the MIUR (Italian Ministry of the University), and 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 Philosophy, University of Pavia.

Lorenzo Magnani  
University of Pavia, Pavia, Italy and Sun Yat-sen University, Guangzhou,  
P.R. China

Ping Li  
Sun Yat-sen University, Guangzhou, P.R. China  
Pavia, Italy/Guangzhou, P.R. China, February 2012

# Contents

## Models, Representation, and Cognition

<b>Scientific Models Are Not Fictions: Model-Based Science as Epistemic Warfare</b> . . . . .	1
<i>Lorenzo Magnani</i>	
<b>An Examination of the Thesis of Models as Representations</b> . . . . .	39
<i>Dachao Li, Ping Li</i>	
<b>On Animal Cognition: Before and After the Beast-Machine Controversy</b> . . . . .	53
<i>Woosuk Park</i>	
<b>From Mindless Modeling to Scientific Models: The Case of Emerging Models</b> . . . . .	75
<i>Tommaso Bertolotti</i>	
<b>The Greenhouse Metaphor and the Greenhouse Effect: A Case Study of a Flawed Analogous Model</b> . . . . .	105
<i>Xiang Chen</i>	
<b>A Study of Model and Representation Based on a Duhemian Thesis</b> . . . .	115
<i>Chuang Liu</i>	
<b>From the Received View to the Model-Theoretic Approach</b> . . . . .	143
<i>Leilei Qi, Huaxia Zhang</i>	
<b>Abduction, Reasoning, and Cognition</b>	
<b>Cognitive Chance Discovery: From Abduction to Affordance</b> . . . . .	155
<i>Akinori Abe</i>	



<b>A Proposal on Belief, Abduction and Interpretation</b> . . . . .	173
<i>Claudio Pizzi</i>	
<b>Not by Luck Alone: The Importance of Chance-Seeking and Silent Knowledge in Abductive Cognition</b> . . . . .	185
<i>Emanuele Bardone</i>	
<b>Cognitive Abduction and the Study of Visual Culture</b> . . . . .	205
<i>María G. Navarro, Noemi de Haro García</i>	
<b>Understanding Scientific Inference in the Natural Sciences Based on Abductive Inference Strategies</b> . . . . .	221
<i>Jun-young Oh</i>	
<b>Moral Intuitions vs. Moral Reasoning. A Philosophical Analysis of the Explanatory Models Intuitionism Relies On</b> . . . . .	239
<i>Sara Dellantonio, Remo Job</i>	
<b>Evolutionary Tolerance</b> . . . . .	263
<i>Luís Moniz Pereira</i>	

# Scientific Models Are Not Fictions

## Model-Based Science as Epistemic Warfare

Lorenzo Magnani

I seem to discern the firm belief that in [natural] philosophizing one must support oneself upon the opinion of some celebrated author, as if our minds ought to remain completely sterile and barren unless wedded to the reasoning of some other person. Possibly he [Lothario Sarsi] thinks that [natural] philosophy is a book of fiction by some writer, like the Iliad or Orlando Furioso, productions in which the least important thing is whether what is written there is true. Well, Sarsi, that is not how matters stand. [Natural] Philosophy is written in this grand book, the universe, which stands continually open to our gaze. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth.

Galileo Galilei, *The Assayer*

**Abstract.** In the current epistemological debate scientific models are not only considered as useful devices for explaining facts or discovering new entities, laws, and theories, but also rubricated under various new labels: from the classical ones, as abstract entities and idealizations, to the more recent, as fictions, surrogates, credible worlds, missing systems, make-believe, parables, functional, epistemic actions, revealing capacities. The paper discusses these approaches showing some of their epistemological inadequacies, also taking advantage of recent results in cognitive science. The main aim is to revise and criticize fictionalism, also reframing the received idea of abstractness and ideality of models with the help of recent results coming from the area of distributed cognition (common coding) and abductive cognition (manipulative). The article also illustrates how scientific modeling activity can be better described taking advantage of the concept of “epistemic warfare”, which sees scientific enterprise as a complicated struggle for rational knowledge in which it is crucial to distinguish epistemic (for example scientific models) from non epistemic (for example fictions, falsities, propaganda) weapons. Finally I will illustrate that it is misleading to analyze models in science by adopting a confounding

---

Lorenzo Magnani

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

Department of Philosophy, Sun Yat-sen University, Guangzhou, P.R. China  
e-mail: [lmagnani@unipv.it](mailto:lmagnani@unipv.it)

mixture of static and dynamic aspects of the scientific enterprise. Scientific models in a static perspective (for example when inserted in a textbook) certainly appear fictional to the epistemologist, but their fictional character disappears in case a dynamic perspective is adopted. A reference to the originative role of thought experiment in Galileo's discoveries and to usefulness of Feyerabend's counterinduction in criticizing the role of resemblance in model-based cognition is also provided, to further corroborate the thesis indicated by the article title.

## 1 Introduction

Current epistemological analysis of the role models in science is often philosophically unproblematic and misleading. Scientific models are now not only considered as useful ways for explaining facts or discovering new entities, laws, and theories, but are also rubricated under various new labels: from the classical ones, abstract entities [Giere, 1988; Giere, 2009; Giere, 2007] and idealizations [Portides, 2007; Weisberg, 2007; Mizrahi, 2011], to the more recent, fictions [Fine, 2009; Woods, 2010; Woods and Rosales, 2010b; Contessa, 2010; Frigg, 2010a; Frigg, 2010b; Frigg, 2010c; Godfrey-Smith, 2006; Godfrey-Smith, 2009; Woods and Rosales, 2010a; Suárez, 2009a; Suárez, 2010], surrogates [Contessa, 2007], credible worlds [Sugden, 2000; Sugden, 2009; Kuorikoski and Lehtinen, 2009], missing systems [Mäki, 2009; Thomson-Jones, 2010], as make-believe [Frigg, 2010a; Frigg, 2010b; Frigg, 2010c; Toon, 2010], parables [Cartwright, 2009b], as functional [Chakravartty, 2010], as epistemic actions [Magnani, 2004a; Magnani, 2004b], as revealing capacities [Cartwright, 2009a]. This proliferation of explanatory metaphors is amazing, if we consider the huge quantity of knowledge on scientific models that had already been produced both in epistemology and in cognitive science. Some of the authors mentioned above are also engaged in a controversy about the legitimacy especially of speaking of fictions in the case of scientific models.

Even if the above studies related to fictionalism have increased knowledge about some aspects of the role of models in science, I am convinced that sometimes they have also generated some philosophical confusion and it seems to me correct (following the suggestion embedded in the title of a recent paper) "to keep quiet on the ontology of models" [French, 2010], and also to adopt a more skeptical theoretical attitude. I think that, for example, models can be considered fictions or surrogates, but this just coincides with a common sense view, which appears to be philosophically empty or, at least, delusory. Models are used in a variety of ways in scientific practice, they can also work as mediators between theory and experiment [Portides, 2007], as pedagogical devices, for testing hypotheses, or for explanatory functions [Bokulich, 2011], but these last roles of models in science are relatively well-known and weakly disputed in the epistemological literature. In this paper I will concentrate on scientific models in creative abductive cognitive processes, which I still consider the central problem of current epistemological research [Hintikka, 1998].

I think that models, both in scientific reasoning and in human perception, are neither mere fictions, simple surrogates or make-believe, nor they are unproblematic idealizations; in particular, models are never *abstract*, contrarily to the received view.<sup>1</sup> In the meantime I aim at substantiating my critique to fictionalism also outlining the first features of my own approach to the role of scientific models in terms of what I call “epistemic warfare”, which sees scientific enterprise as a complicated struggle for rational knowledge in which it is crucial to distinguish epistemic (for example scientific models) from non epistemic (for example fictions, falsities, propaganda, etc.) weapons.<sup>2</sup> I certainly consider scientific enterprise a complicated epistemic warfare, so that we could plausibly expect to find fictions in this struggle for rational knowledge. Are not fictions typical of any struggle which characterizes the conflict of human coalitions of any kind? During the Seventies of the last century Feyerabend [Feyerabend, 1975] clearly stressed how, despite their eventual success, the scientist’s claims are often far from being evenly proved, and accompanied by “propaganda [and] psychological tricks in addition to whatever intellectual reasons he has to offer” (p. 65), like in the case of Galileo. These tricks are very useful and efficient, but one count is the *epistemic* role of reasons scientist takes advantage of, such as scientific models, which for example directly govern the path to provide a new intelligibility of the target systems at hand, another count is the *extra-epistemic* role of propaganda and rhetoric, which only plays a mere – positive or negative – ancillary role in the epistemic warfare. So to say, these last aspects support scientific reasoning providing non-epistemic weapons able for example to persuade other scientists belonging to a rival “coalition” or to build and strengthen the coalition in question, which supports a specific research program, for example to get funds.

I am neither denying that models as idealizations and abstractions are a pervasive and permanent feature of science, nor that models, which are produced with the aim of finding the consequences of theories – often very smart and creative – are very important. I just stress that the “fundamental” role played by models in science is the one we find in the core conceptual discovery processes, and that these kinds of models cannot be indicated as fictional at all, because they are *constitutive* of new scientific frameworks and new empirical domains.<sup>3</sup>

---

<sup>1</sup> This does not mean that the standard epistemological concept of abstract model is devoid of sense, but that it has to be considered in a Pickwickian sense.

<sup>2</sup> The characteristic feature of *epistemic* weapons is that they are value-directed to the aim of promoting the attainment of scientific truth, for example through predictive and empirical accuracy, simplicity, testability, consistency, etc.: in this perspective I basically agree with the distinction between epistemic and non-epistemic values as limpidly depicted in [Steel, 2010].

<sup>3</sup> In this last sense the capacity of scientific models to constitute new empirical domains and so new *knowability* is ideally related to the emphasis that epistemology, in the last century, put on the theory-ladenness of scientific facts (Hanson, Popper, Lakatos, Kuhn): in this light, the formulation of observation statements presupposes significant knowledge, and the search for new observability in science is guided by scientific modeling. On this issue cf. also [Bertolotti, 2012], this volume.

Suárez [Suárez, 2009a] provides some case studies, especially from astrophysics and concerning quantum model of measurement, emphasizing the inferential function of the supposed to be “fictional” assumptions in models: I deem this function to be ancillary in science, even if often highly innovative. Speaking of the Thomson’s plum pudding model Suárez maintains that, basically “The model served an essential pragmatic purpose in generating quick and expedient inference at the theoretical level, and then in turn from the theoretical to the experimental level. It articulated a space of reasons, a background of assumptions against which the participants in the debates could sustain their arguments for and against these three hypotheses” (p. 163). In these cases the fact that various assumptions of the models are empirically false is pretty clear and so is the “improvement in the expediency of the inferences that can be drawn from the models to the observable quantities” (p. 165):<sup>4</sup> the problem is that in these cases models, however, are not fictions – at least in the minimal unequivocal sense of the word as it is adopted in the literary/narrative frameworks – but just the usual idealizations or abstractions, already well-known and well studied, as devices, stratagems, and strategies that lead to efficient results and that are not discarded just because they are not fake chances from the perspective of scientific rationality.<sup>5</sup> Two consequences derive:

- the role of models as “expediency of the inferences” in peripheral aspects of scientific research, well-known from centuries in science, does not have to be confused with the *constitutive* role of modeling in the central creative processes, when new conceptually revolutionary perspectives are advanced;
- models are – so to say – just models that idealize and/or abstract, but these last two aspects have to be strictly criticized in the light of recent epistemologico/cognitive literature as special kinds of epistemic actions, as I will illustrate in section 3 below: abstractness and ideality cannot be solely related to empirical inadequacy and/or to theoretical incoherence [Suárez, 2009a, p. 168], in a static view of the scientific enterprise.

In sum, I will illustrate that there is no need of reframing – in the new complicated and intellectualistic lexicon of fictions (and of the related metaphors) – what is already well-known thanks to the tradition of philosophy of science. We have to remorselessly come back to Newton’s famous motto “hypotheses non fingo”, which has characterized for centuries the spirit of modern science: “I have not as yet been able to discover the reason for these properties of gravity from phenomena, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be

<sup>4</sup> It has to be added that Suárez does not certainly conflate scientific modeling with literary fictionalizing. He clearly distinguishes scientific fictions from other kinds of fictions – the scientific ones are constrained by both the logic of inference and, in particular, the requirement to fit in with the empirical domain [Suárez, 2009a; Suárez, 2010] – in the framework of an envisaged compatibility of “scientific fiction” with realism. This epistemological acknowledgment is not often present in other stronger followers of fictionalism.

<sup>5</sup> I discussed the role of chance-seeking in scientific discovery in [Magnani, 2007]. For a broader discussion on the role of luck and chance-seeking in abductive cognition see also [Bardone, 2011], and [Bardone, 2012], this volume.

called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction” [Newton, 1999, p. 493].

## 2 Models Are Not Fictions. The Inconsistency of the Argument of Imperfect Fit

Should scientific models be regarded as works of fictions? At the beginning of the previous section 1 said that models, both in scientific reasoning and in human perception, are neither mere fictions, simple surrogates or make-believe, nor they are unproblematic idealizations; in particular, models are never abstract, contrarily to the received view. Let us outline in this section the first problem, related to the *fictionalist* nature of models. I will return to this problem in section 3, in which also the problem of the *abstractness* of models will be deeply illustrated: as for now we can note that, in a philosophical naturalistic framework, where all phenomena and thus also cognition, gain a fundamental eco-physical significance, models are always material objects, either when we are dealing with concrete diagrams, physical or computational models, or when we face human “mental models”, which at the end “are” particular, unrepeatable, but ever-changing configurations and transformations of neural networks and chemical distributions at the level of human brains. Indeed, defending in this paper an interdisciplinary approach we are simply re-engaged in one of the basic tenets of the philosophical mentality, now enriched by a naturalistic commitment, which acknowledges the relevance of scientific results of cognitive research.

If, ontologically, models are imaginary objects in the way objects of fictions are imaginary objects, I cannot see them as situated in any “location” different from the brain, so that they are imaginary in so far as they are just “mental” models. Like Giere contends:

In spite of sharing an ontology as imagined objects, scientific models and works of fiction function in different cultural worlds. One indication of this difference is that, while works of fiction are typically a product of a single author’s imagination, scientific models are typically the product of a collective effort. Scientists share preliminary descriptions of their models with colleagues near and far, and this sharing often leads to smaller or larger changes in the descriptions. The descriptions, then, are from the beginning intended to be public objects. Of course, authors of fiction may share their manuscripts with family and colleagues, but this is not part of the ethos of producing fiction. An author would not be professionally criticized for delivering an otherwise unread manuscript an editor. Scientists who keep everything to themselves before submitting a manuscript for publication are regarded as peculiar and may be criticized for being excessively secretive [Giere, 2009, p. 251].

Moreover, to consider models as fictions would destroy the well regarded distinction between science and science fiction. This attitude can present cultural dangers: is science just a matter of fictions? Both kinds of fictions (scientific and literary)

certainly provide insights on something “real”, that is they aim at *representing* aspects of the world (for example *War and Peace*, Giere says, provides insight into the “human condition”) but often various genres of literary fictions are simply finalized to entertain. Even if both contain imaginary objects, the processes that govern their formation and what from them is derived are very dissimilar, as I will further describe in section 3. Representation in science is always related to criteria of scope, accuracy, precision and detail – Giere says – and further notes: “Remember the many models that were proposed and rejected in the race for the double helix because they failed adequately to represent the structure of DNA molecules. In the realm of fantasy, such criticisms are not appropriate. It is no criticism of the Harry Potter novels that there is no community of genuine wizards. Nor is it a criticism of *War and Peace* that its main characters did not exist” [Giere, 2009, p. 252]. The fact that a scientific model, relating to the “real” world, seems to be a fiction – that is to say, the fact it does not perfectly fit to any real system – does not authorize us to regard the overall model as a work of fiction, because it does not function like a work of fiction such as novels or so.

Finally, I strongly agree with Giere that “In fact, the argument from imperfect fit to a functionally fictional status for models proves far too much” [Giere, 2009, p. 254], because it is typical of every cognition the involvement of ideal categorization and schematization, so that most of what everyone thinks and perceives should be regarded as fictional:<sup>6</sup>

It seems to me that the assimilation of scientific models to works of fiction presupposes an exaggerated conception of nonfiction. On this conception, a genuine work of nonfiction has to provide “the truth, the whole truth, and nothing but the truth”. Thus, the realization that scientists are mostly in the business of constructing models that never provide a perfect fit to the world leads to the unwarranted conclusion that scientists are in the business of producing fictional accounts of the world (cit.)

The problem is that models help reach success in experimental outcomes, because they instead fit to designated aspects of the world:

[...] the view that scientific models are ontologically like works of fiction in being imaginary creations not only does not uniquely support fictionalism, but is compatible with a moderate realism. There is nothing in this notion of a scientific model that prevents identifying elements of models with things traditionally classified as “unobservable”. On the other hand, as discussed earlier in this chapter, some elements of models may not be identified with anything in the world (cit., p. 256).

---

<sup>6</sup> Mizrahi [Mizrahi, 2011] seems to support – in the linguistic perspective about the role of “facticity” in scientific cognition – a similar point of view about the coherence of seeing scientific “idealized” models as “quasi-factive”: “[...] if [scientific] understanding is (quasi) factive, then we can attribute this sort of cognitive success to scientists when they employ idealizations, such as the Ideal Gas Law, precisely because they mirror the facts to some extent. That is to say, in the case of the Ideal Gas Law, it is precisely because of the agreement between the predictions of the gas laws and the behavior of gases (under specified conditions of temperature and pressure) that we attribute cognitive success to scientists in this case. Otherwise, it seems, we would say that scientists don’t understand the behavior of gases at all”.

I confess that I would not encourage epistemologists to engage in debates about “realism” against “fictionalism”, or about problems like “is fictionalism compatible with realism?”, etc. [Suárez, 2010], because the adoption of these old pre-Kantian categories is in my opinion philosophically sterile. After all, the same discussions about a privileged *level* of reality (able to demarcate everything else, for example “fictions”) could be easily substituted by an equally coherent view about the consistency of various *levels* of reality, where the referents of fictions could be easily included.

It is not that “fictions provide inferential shortcuts in models; and the fact that this is the main or only reason for their use distinguishes them as fictional” [Suárez, 2010, p. 239], even if Vaihinger would agree with this functionalist perspective on fictions.<sup>7</sup> Indeed, even if it is not decisive to say “that the inferential characterisation provides a way to distinguish precisely scientific from non-scientific uses of fiction”, models used in non-scientific practices may also trigger inferences, and the problem here is more fundamental. In science, models are not used and intended as fictions, they are just labeled as fictions because of a juxtaposition of some recent philosophers of science, who certainly in this way render the scientific enterprise more similar to other more common modes of human cognition: after all fictions are ubiquitous in human cognition, and science is a cognitive activity like others. Unfortunately science never aimed to provide “fictions” at the basic levels of its activities, so that the recent fictionalism does not add new and fresh knowledge about the status of models in science, and tends to obfuscate the distinctions between different areas of human cognition, such as science, religion, arts, and philosophy. In the end, “epistemic fictionalism” tends to enforce a kind “epistemic concealment”, which can obliterate the actual gnoseological finalities of science, shading in a kind of debate about entities and their classification that could remind of medieval scholasticism.

### 3 Models Are Distributed and Never Abstracts: Model-Based Science as Epistemic Warfare

At the beginning of the previous section I advanced the hypothesis that models, both in scientific reasoning and in human perception, are neither mere fictions, simple surrogates or make-believe, nor they are unproblematic idealizations, and I also specifically contended that models are never *abstract* or *ideal*, contrarily to the received view: they do not live – so to say – in a kind of mysterious Popperian *World*

---

<sup>7</sup> Suárez’s approach to scientific models as fictions is actually more sophisticated than it may appear from my few notes. Basically, Suárez does not defend the view according to which models are fictions: even if he defends the view that models contain or lead to fictional assumptions, he explicitly rejects the identification of models and fictions, preferring instead to stay “quietist” about the ontology of models, and focusing rather on modeling as an activity – see in particular his introduction to the 2009 Routledge volume he edited entitled *Fictions in Science* [Suárez, 2009b].



3. Let us deepen this second problem concerning the abstract and ideal nature of models in scientific reasoning.

First of all, within science the adopted models are certainly constructed on the basis of multiple constraints relating to the abstract laws, principles, and concepts, when clearly available at a certain moment of the development of a scientific discipline. At the same time we have to immediately stress that the same models are always *distributed* material entities, either when we are dealing with concrete diagrams or physical and computational models, or when we face human “mental models”, which at the end are indeed particular, unrepeatable, and ever-changing configurations and transformations of neural networks and chemical distributions at the level of human brains. In this perspective we can say that models are “abstract” only in a Pickwickian sense, that is as “mental models”, shared to different extents by groups of scientists, depending on the type of research community at stake. This cognitive perspective can therefore help us in getting rid of the ambiguities sparked by the notion of abstractness of models.

I contend that the so-called *abstract model* can be better described in terms of what Nersessian and Chandrasekharan [Nersessian and Chandrasekharan, 2009] call *manifest model*: when the scientific collective decides whether the model is worth pursuing, and whether it would address the problems and concepts researchers are faced with, it is an internal model and it is manifest because it is shared and “[. . .] allows group members to perform manipulations and thus form common movement representations of the proposed concept. The manifest model also improves group dynamics” [Chandrasekharan, 2009, p. 1079]. Of course the internal representation presents slight differences in each individual’s brain, but this does not impede that the various specific representations are clearly thought to be “abstract” insofar as they are at the same time “conceived” as referring to a unique model. This model, at a specific time, is considered “manifest”, in an atmosphere of common understanding. Nevertheless, *new* insights/modifications in the internal manifest model usually occur at the individual level, even if the approach to solve a determinate problem through the model at stake is normally shared by a specific scientific collective: the singular change can lead to the solution of the problems regarding the target system and so foster new understanding. However, new insights/modifications can also lead to discard the model at stake and to build another one, which is expected to be more fruitful and which possibly can become the new manifest model. Moreover, some shared manifest models can reach a kind of stability across the centuries and the scientific and didactic communities, like in the case of the ideal pendulum, so that they optimally reverberate the idea of high *abstractness* of scientific models.

If we comply with a conception of the mind as “extended”, we can say that the mind’s guesses – both instinctual and reasoned – can be classified as plausible hypotheses about “nature” because the mind grows up *together with* the representational delegations to the external world that mind itself has made throughout the

history of culture by constructing the so-called cognitive niches.<sup>8</sup> Consequently, as I have already anticipated few lines above, not only scientific models are never abstracts/ideal, they are always distributed. Indeed, in the perspective of distributed (and embodied) cognition [Hutchins, 1999] a recent experimental cognitive research [Chandrasekharan, 2009] further provides deep and fresh epistemological insight into the old problem of abstractness and ideality of models in scientific reasoning. The research illustrates two concrete external models, as functional and behavioral approximations of neurons, one physical (in-vitro networks of cultured neurons) and the other consisting in a computational counterpart, as recently built and applied in a neural engineering laboratory.<sup>9</sup> These models are clearly recognized as external systems – external artifacts more or less intentionally<sup>10</sup> prepared, exactly like concrete diagrams in the case of ancient geometry – interacting with the internal corresponding models of the researchers, and they aim at generating new concepts and control structures regarding target systems.

The external models in general offer more plasticity than the internal ones and lower memory and cognitive load for the scientist's minds. They also incorporate constraints imposed by the medium at hand that also depend on the intrinsic and immanent cognitive/semiotic delegations (and the relative established conventionality) performed by the model builder(s): artificial languages, proofs, new figures, examples, computational simulations, etc.<sup>11</sup> It is obvious that the information (about model behavior) from models to scientists flow through perception (and not only through visualization as a mere representation – as we will see below, in the case of common coding also through “movements in the visualization are also a way of generating equivalent movements in body coordinates” [Chandrasekharan, 2009, p. 1076]).

Perception persists in being the vehicle of model-based and motor information to the brain. We see at work that same perception that Peirce speculatively analyzed as that complicated philosophical structure I illustrated in my book on abductive cognition.<sup>12</sup> Peirce explains to us that some basic human model-based ways of knowing, that is *perceptions*, are abductions, and thus that they are hypothetical

<sup>8</sup> The concept of cognitive niche is illustrated in detail in [Odling-Smee *et al.*, 2003].

<sup>9</sup> An analysis of the differences between models in biology and physics and of the distinction between natural, concrete, and abstract models is illustrated in [Rowbottom, 2009]; unfortunately, the author offers a description of abstract models that seems to me puzzling, and falls under the criticism I am illustrating in the present paper.

<sup>10</sup> I have to note that manipulative abduction – see below subsection 3.1 – also happens when we are *thinking through doing* (and not only, in a pragmatic sense, about doing). This kind of action-based cognition can hardly be intended as completely intentional and conscious.

<sup>11</sup> On the cognitive delegations to external artifacts see [Magnani, 2009, chapter three, section 3.6]. A useful description of how formats also matter in the case of external hypothetical models and representations, and of how they provide different affordances and inferential chances, cf. [Vorms, 2010].

<sup>12</sup> The complicated analysis of some seminal Peircean philosophical considerations concerning abduction, perception, inference, and instinct, which I consider are still important to current cognitive and epistemological research, is provided in [Magnani, 2009, chapter five].

and withdrawable. Moreover, given the fact that judgments in perception are fallible but indubitable abductions, we are not in any psychological condition to conceive that they are false, as they are unconscious habits of inference. Hence, these fundamental – even if non scientific – model-based ways of cognizing are constitutively intertwined with inferential processes. *Unconscious* cognition legitimately enters these processes (and not only in the case of some aspects of perception – remind the process, in scientific modeling, of “thinking through doing”, I have just quoted above in footnote 10), so that model-based cognition is typically performed in an unintentional way. The same happens in the case of emotions, which provide a quick – even if often highly unreliable – abductive appraisal/explanation of given data, which is usually anomalous or inconsistent. It seems that, still in the light of the recent results in cognitive science I have just described, the importance of the model-based character of perception stressed by Peirce is intact. This suggests that we can hypothesize a continuum from construction of models that actually *emerge* at the stage of perception, where models are operating with the spontaneous application of abductive processes to the high-level model activities of more or less intentional modelers ([Park, 2011], and [Bertolotti, 2012], this volume), such as scientists.<sup>13</sup> Finally, if perception cannot be wrong, given the fact that judgments in perception are fallible but indubitable abductions, as I have just illustrated, then these judgments should not be regarded as *fictional*.

### ***3.1 Perception-Action Common Coding as an Example of “On-Line” Manipulative Abduction***

The cognitive mechanism carefully exploited and illustrated in [Chandrasekharan, 2009] takes advantage of the notion of *common coding*,<sup>14</sup> recently studied in cognitive science and closely related to embodied cognition, as a way of explaining the special kind of “internal-external coupling”, where brain is considered a control mechanism that coordinates action and movements in the world. Common coding hypothesizes

---

<sup>13</sup> On the puzzling problem of the “modal” and “amodal” character of the human brain processing of perceptual information, and the asseveration of the importance of grounded cognition, cf. [Barsalou, 2008a; Barsalou, 2008b].

<sup>14</sup> “The basic argument for common coding is an adaptive one, where organisms are considered to be fundamentally action systems. In this view, sensory and cognitive systems evolved to support action, and they are therefore dynamically coupled to action systems in ways that help organisms act quickly and appropriately. Common coding, and the resultant replication of external movements in body coordinates, provides one form of highly efficient coupling. Since both biological and nonbiological movements are equally important to the organism, and the two movements interact in unpredictable ways, it is beneficial to replicate both types of movements in body coordinates, so that efficient responses can be generated” [Chandrasekharan, 2009, p. 1069]: in this quoted paper the reader can find a rich reference to the recent literature on embodied cognition and common coding.

[...] that the execution, perception, and imagination of movements share a common representation (coding) in the brain. This coding leads to any one of these three (say perception of an external movement), automatically triggering the other two (imagination and execution of movement). One effect of this mechanism is that it allows any perceived external movement to be instantaneously replicated in body coordinates, generating a dynamic movement trace that can be used to generate an action response. The trace can also be used later for cognitive operations involving movement (action simulations). In this view, movement crosses the internal/external boundary *as movement*, and thus movement could be seen as a “lingua franca” that is shared across internal and external models, if both have movement components, as they tend to do in science and engineering [Chandrasekharan, 2009, p. 1061].

Common coding refers to a representationalist account, but representation supports a motor simulation mechanism “which can be activated across different timescales – instantaneous simulation of external movement, and also extended simulations of movement. The latter could be online, that is, linked to an external movement (as in mental rotations while playing Tetris, see [Kirsh and Maglio1994]), or can be offline (as in purely imagined mental rotation)” [Chandrasekharan, 2009, p. 1072]. Furthermore

1. given the fact models in science and engineering often characterize phenomena in terms of bodies and particles, motor simulations are important to understand them, and the lingua franca guarantees integration between internal and external models;
2. the manipulation of the external models creates new patterns that are offered through perception to the researchers (and across the whole team, to possibly reach that shared “manifest model” I have illustrated above), and “perturbs” (through experimentation on the model that can be either intended or random) their movement-based internal models possibly leading “[...] to the generation of nonstandard, but plausible, movement patterns in internal models, which, in combination with mathematical and logical reasoning, leads to novel concepts” (cit., p. 1062);
3. this hybrid combination with mathematical and logical reasoning, and possible other available representational resources stored in the brain, offers an example of the so-called multimodality of abduction.<sup>15</sup> Not only both data and theoretical adopted hypotheses, but also the intermediate steps between them – i.e. for example, models – can have a full range of verbal and sensory representations, involving words, sights, images, smells, etc. and also kinesthetic and motor experiences and feelings such as satisfaction, and thus all sensory modalities. Furthermore, each of these cognitive levels – for example the mathematical ones, often thought as presumptively *abstract* [does this authorize us to say they are fictional?] – actually consist in intertwined and flexible models (*external* and *internal*) that can be analogically referred to the Peircean concept of the “compound conventional sign”, where for example sentential and logical aspects coexist with model-based features. For Peirce, *iconicity hybridates log-*

---

<sup>15</sup> On the concept of multimodal abduction see chapter four of [Magnani, 2009].

*icality*: the sentential aspects of symbolic disciplines like logic or algebra coexist with model-based features – iconic. Indeed, sentential features like symbols and conventional rules<sup>16</sup> are intertwined with the spatial configuration, like in the case of “compound conventional signs”. Model-based iconicity is always present in human reasoning, even if often hidden and implicit;<sup>17</sup>

4. it is the perturbation I have described above that furnishes a chance for change, often innovative, in the internal model (new brain areas can be activated creating new connections, which in turn can motivate further manipulations and revisions of the external model): it is at this level that we found the scientific cognitive counterpart of what has been always called in the tradition of philosophy and history of science, scientific imagination.<sup>18</sup>

It is worth to note that, among the advantages offered by the external models in their role of perturbing the internal ones, there are not only the unexpected features that can be offered thanks to their intrinsic materiality, but also more neutral but fruitful devices, which can be for example exemplified thanks to the case of externalized mathematical symbols: “Apparently the brain immediately translates a positive integer into a mental representation of its quantity. By contrast, symbols that represent non-intuitive concepts remain partially semantically inaccessible to us, we do not reconstruct them, but use them as they stand” [De Cruz and De Smedt, 2011]. For example, it is well-known that Leibniz adopted the notation  $dx$  for the infinitesimals he genially introduced, and called them *fictions bien fondées*, given their semantic paradoxical character: they lacked a referent in Leibnizian infinitesimal calculus, but

<sup>16</sup> Written natural languages are intertwined with iconic aspects too. Stjernfelt [Stjernfelt, 2007] provides a full analysis of the role of icons and diagrams in Peircean philosophical and semiotic approach, also taking into account the Husserlian tradition of phenomenology.

<sup>17</sup> It is from this perspective that [sentential] syllogism and [model-based] perception are seen as rigorously intertwined. Consequently, there is no sharp contrast between the idea of cognition as perception and the idea of cognition as something that pertains to logic. Both aspects are inferential in themselves and fruit of sign activity. Taking the Peircean philosophical path we return to observations I always made when speaking of the case of abduction: cognition is basically *multimodal*.

<sup>18</sup> In a perspective that does not take into account the results of cognitive science but instead adopts the narrative/literary framework about models as make-believe, Toon [Toon, 2010] too recognizes the role of models in perturbing mental models to favor imagination: “Without taking a stance in the debate over proper names in fiction, I think we may use Walton’s analysis to provide an account of our prepared description and equation of motion. We saw [...] that these are not straightforward descriptions of the bouncing spring. Nevertheless, I believe, they do represent the spring, in Walton’s sense: they represent the spring by prescribing imaginings about it. When we put forward our prepared description and equation of motion, I think, those who are familiar with the process of theoretical modelling understand that they are to imagine certain things about the bouncing spring. Specifically, they are required to imagine that the bob is a point mass, that the spring exerts a linear restoring force, and so on” (p. 306).

were at the basis of plenty of new astonishing mathematical results.<sup>19</sup> De Cruz and De Smedt call this property of symbols “semantic opacity”, which renders them underdetermined, allowing further creative processes as symbols that can be relatively freely exploited in novel contexts for multiple cognitive aims. Semantic opacity favors a kind of reasoning that is unbiased by the intuitive aspects possibly involving stereotypes or intended uncontrolled interpretations, typical of other external models/representations.

Peirce too was clearly aware, speaking of the model-based aspects of deductive reasoning, that there is an “experimenting upon this image [the external model/diagram] in the imagination”, where the idea that human imagination is always favored by a kind of prosthesis, the external model as an “external imagination”, is pretty clear, even in case of classical geometrical deduction: “[. . .] namely, deduction consists in constructing an icon or diagram the relations of whose parts shall present a complete analogy with those of the parts of the object of reasoning, of experimenting upon this image in the imagination and of observing the result so as to discover unnoticed and hidden relations among the parts” [Peirce, 1931-1958, 3.363].

Analogously, in the case at stake, the computational model of neuronal behavior, by providing new chances in terms of control, visualizations, and costs, is exactly the peculiar tool able to favor manipulations which trigger the new idea of the “spatial activity pattern of the spikes” [Chandrasekharan, 2009, p. 1067].

### 3.2 *Fictions or Epistemic Weapons?*

Thanks to the cognitive research I have illustrated in the previous subsection, we are faced with the cognitive modern awareness of what also implicitly underlies Peircean speculations: nature fecundates the mind because it is through a disembodiment and extension of the mind in nature (that is, so to say, “artificialized”) that in turn nature affects the mind. Models are built by the mind of the scientist(s), who first delegate “meanings” to external artifacts: mind’s “internal” representations are “extended” in the environment, and later on shaped by processes that are occurring through the constraints found in “nature” itself; that is that external nature that consists of the “concrete” model represented by the artifact, in which the resulting aspects and modifications/movements are “picked up” and in turn re-represented in the human brain. It is in this perspective that we can savor, now in a naturalistic framework, the speculative Aristotelian anticipation that “nihil est in intellectu quod prius non fuerit in sensu”. In such a way – that is thanks to the information that flows from the model – the scientists’ internal models are rebuilt and further

---

<sup>19</sup> To confront critiques and suspects about the legitimacy of the new number  $dx$ , Leibniz prudently conceded that  $dx$  can be considered a fiction, but a “well founded” one. The birth of non-standard analysis, an “alternative calculus” invented by Abraham Robinson [Robinson, 1966], based on infinitesimal numbers in the spirit of Leibniz’s method, revealed that infinitesimals are not at all fictions, through an extension of the real numbers system  $\mathbb{R}$  to the system  $\mathbb{R}^*$  containing infinitesimals smaller in the absolute value than any positive real number.

refined and the resulting modifications can easily be seen as guesses – both instinctual and reasoned, depending of the brain areas involved, that is as plausible abductive hypotheses about the external extra-somatic world (the target systems). I repeat, the process can be seen in the perspective of the theory of cognitive niches: the mind grows up together with its representational delegations to the external world that has made itself throughout the history of culture by constructing the so-called cognitive niches. In this case the complex cognitive niche of the scientific lab is an *epistemological* niche, expressly built to increase knowledge following rational methods, where “*people, systems, and environmental affordances*” [Chandrasekharan, 2009, p. 1076] work together in an integrated fashion.

Even if Chandrasekharan and Nersessian’s research deals with models which incorporate movement, and so does not consider models that are not based on it, it provides an useful example able to stress the distributed character of scientific models, and the true type of abstractness/ideality they possess, so refreshing these notions that come from the tradition of philosophy of science. The analysis of models as material, mathematical, and fictional – and as “abstract objects” – provided by Contessa [Contessa, 2010], where “a model is an actual abstract object that stands for one of the many possible concrete objects that fit the generative description of the model” (p. 228) would take advantage of being reframed in the present naturalistic perspective. The same in the case of Frigg [Frigg, 2010c], who contends a fictionalist view and says “Yet, it is important to notice that the model-system is not the same as its [verbal] description; in fact, we can re-describe the same system in many different ways, possibly using different languages. I refer to descriptions of this kind as model-descriptions and the relation they bear to the model-system as *p*-representation” (pp. 257–258). Indeed, Contessa’s reference to models as “actual abstract objects” and Frigg’s reference to models as abstract “model-systems” would take advantage of the cognitive perspective I am presenting here: where are they located, from a naturalistic point of view? Are they mental models? If they are mental models, like I contend, this should be more clearly acknowledged.

Hence, in my perspective models cannot be considered neither abstract (in the traditional ambiguous sense) nor fictional: scientist do not have any intention to propose fictions, instead they provide models as tools that reshape a generic cognitive niche as an epistemological niche to the aim of performing a genuine struggle for representing the external world. Models, the war machines used in this struggle, which I call *epistemic warfare*, to stress the determined – strictly epistemic – dynamism of the adopted tools that are at stake, are not illusional fictions or stratagems used for example to cheat nature or swindle human beings, but just concrete, unambiguous, and well disposed tactical intermediate weapons able to strategically “attack” nature (the target systems) to further unveil its structure. Contrarily, fictions in works of fictions are for example meant to unveil human life and characters in new esthetic perspectives and/or to criticize them through a moral teaching, while fictions and stratagems in wars are meant to trick the enemy and possibly destroy the eco-human targets.

I contend that epistemologists do not have to forget that various cognitive processes present a “military” nature, even if it is not evident in various aspects and

uses of syntactilized human natural language and in abstract knowledge.<sup>20</sup> It is hard to directly see this “military intelligence”<sup>21</sup> in the many *epistemic* functions of natural language, for example when it is simply employed to transmit scientific results in an academic laboratory situation, or when we gather information from the Internet – expressed in linguistic terms and numbers – about the weather. However, we cannot forget that even the more abstract character of knowledge packages embedded in certain uses of language (and in hybrid languages, like in the case of mathematics, which involves considerable symbolic parts) still plays a significant role in changing the moral behavior of human collectives. For example, the production and the transmission of new scientific knowledge in human social groups not only operates on information but also implements and distributes roles, capacities, constraints and possibilities of actions. This process is intrinsically moral because in turn it generates precise distinctions, powers, duties, and chances which can create new between-groups and in-group violent (often) conflicts, or reshape older pre-existent ones.

New theoretical biomedical knowledge about pregnancy and fetuses usually has two contrasting moral/social effects, 1) a better social and medical management of childbirth and related diseases; 2) the potential extension or modification of conflicts surrounding the legitimacy of abortion. In sum, even very abstract bodies of knowledge and more innocent pieces of information enter the semio/social process which governs the identity of groups and their aggressive potential as coalitions: deductive reasoning and declarative knowledge are far from being exempt from being accompanied by argumentative, deontological, rhetorical, and dialectic aspects. For example, it is hard to distinguish, in an eco-cognitive setting, between a kind of “pure” (for example deductive) inferential function of language and an argumentative or deontological one. For example, the first one can obviously play an associated argumentative role. However, it is in the arguments traditionally recognized as fallacious, that we can more clearly grasp the military nature of human language and especially of some hypotheses reached through fallacies.

Hence, we have to be aware that science imposes itself as a paradigm of producing knowledge in a certain “decent” way, but at the same time it *de facto* belongs to the cross-disciplinary warfare that characterizes modernity: science more or less conflicts with other non scientific disciplines, religions, literature, magic, etc., and also implicitly orders and norms societies through technological products which impose behaviors and moral conducts. Of course scientific cognitive processes – *sensu strictu*, inside scientific groups as coalitions – also involve propaganda, like Feyerabend says, for instance to convince colleagues about a hypothesis or a method, but propaganda is also externally addressed to other private and public coalitions and common people, for example to get funds (a fundamental issue often disregarded

---

<sup>20</sup> I extendedly treated the relationship between cognition and violence in my [Magnani, 2011].

<sup>21</sup> I am deriving this expression from René Thom [Thom, 1988], who relates “military intelligence” to the role played by language and cognition in the so-called *coalition enforcement*, that is at the level of their complementary effects in the affirmation of moralities and related conducts, and the consequent perpetration of possible violent punishments.



in the contemporary science is the cost of producing new models) or to persuade about the value of scientific knowledge. Nevertheless the core cognitive process of science is based on avoiding fictional and rhetorical devices when the production of its own regimen of truth is at stake. Finally, science is exactly that enterprise which produces truths that establish themselves as the paradigms for demarcating fictions and so “irrational” or “arational” ways of knowing.

I am aware of the fact that epistemological fictionalism does not consider fictions forgery or fake, that is something “far from being execrable”, instead, something “we cherish” [Frigg, 2010c, p. 249], but to say that scientific and literary fictions are both “good” fictions is a bit of a theoretical oversemplification, because it is science that created, beyond literature and poetry, *new* kinds of models committed to a specific production of truth, constitutively aiming at not being fictional.<sup>22</sup> I confess I cannot see how we can speak of the ideal pendulum in the same way we speak of Anna Karenina: it seems to me that we are running the risk of inadvertently opening the gates of epistemology to a kind of relativistic post-modernism *à la mode*, even if fictionalists seem to avoid this possible confusion by producing – often useful – taxonomies about the slight differences between fictions in science and in other cognitive practices.

In overall, I am convinced that introducing the word fiction in epistemology adds a modest improvement to the analysis of topics like inference, explanation, creativity, etc., but just an attractive new lexicon, which takes advantage of some seductive ideas coming for example from the theory of literary fictions. Anna Karenina and the in-vitro model<sup>23</sup> are very different. In actual scientific practice, a model becomes fictional only *after* the community of researchers has recognized it as such, *because* it has *failed* in fruitfully representing the target systems. In these cases a model is simply discarded. Tolstoy might have discarded the character of Anna Karenina as an inappropriate fiction for some contemporary esthetic purpose (for instance, had she failed, in her author’s opinion, to veraciously represent a female member of Russia’s high society at the end of XIX century), but he would have substituted her with yet another – just as fictional – character, doomed to *remain* fictional for ever.<sup>24</sup>

As I already said, conversely a scientific model is recognized as fictional in a cognitive (often creative) process when it is assessed to be unfruitful, by applying a kind of *negation as failure* [Clark, 1978; Magnani, 2001a]: it becomes fictional in the mere sense that it is falsified (even if “weakly” falsified, by failure).<sup>25</sup> Methodologically, negation as failure is a process of elimination that parallels what

<sup>22</sup> Cf. below, subsection 3.6.

<sup>23</sup> Indeed, in the recent epistemological debate about fictions, even the whole “experimental systems” are reframed as “materialized fictional ‘worlds’” [Rouse, 2009, p. 51].

<sup>24</sup> Giere usefully notes that “Tolstoy did not intend to represent actual people except in general terms” and that, on the contrary, a “primary function [of models in science], of course, is to represent physical processes in the real world” [Giere, 2007, p. 279].

<sup>25</sup> On the powerful and unifying analysis of inter-theory relationships, which involves the problem of misrepresenting models – and their substitution/adjustment – and of incompleteness of scientific representation, in terms of partial structural similarity, cf. [Bueno and French, 2011] and the classic [da Costa and French, 2003].

Freud describes in the case of constructions (the narratives the analyst builds about patient's past psychic life) abandoned because they do not help to proceed in the therapeutic psychoanalytic process: if the patient does not provide new "material" which extends the proposed construction, "if", as Freud declares, "[...] nothing further develops we may conclude that we have made a mistake and we shall admit as much to the patient at some suitable opportunity without sacrificing any of our authority". The "opportunity" of rejecting the proposed construction "will arise" just "[...] when some new material has come to light which allows us to make a better construction and so to correct our error. In this way the false construction drops out, as if it has never been made; and indeed, we often get an impression as though, to borrow the words of Polonius, our bait of falsehood had taken a carp of truth" [Freud, 1953-1974, vol. 23, 1937, p. 262].

Similarly, for example in a scientific discovery process, the scientific model is simply eliminated and labeled as "false", because "new material has come to light" to provide a better model which in turn will lead to a new knowledge that supersedes or refines the previous one, and so the old model is buried in the necropolis of the unfruitful/dead models. Still, similarly, in the whole scientific enterprise, also a successful scientific model is sometimes simply eliminated (for example the ether model) together with the theory to which that model belonged, and so the old model is buried in yet another necropolis, that of the abandoned "historical" models, and yes, in this case, it can be plausibly relabeled as a fiction.

A conclusion in tune with my contention against the fictional character of scientific models is reached by Woods and Rosales [Woods and Rosales, 2010a], who offer a deep and compelling logico-philosophical analysis of the problem at stake. They contend that it is extremely puzzling to extend the theory of literary and artistic fictions to science and other areas of cognition. Whatever we say of the fictions of mathematics and science, there is "nothing true of them in virtue of which they are *literary fictions*" (p. 375). They correctly note that "Saying that scientific stipulation is subject to normative constraints is already saying something quite different from what should be said about literary stipulation":

We also see that scientific stipulation is subject to a *sufferance* constraint, and with it to factors of timely goodness. A scientist is free to insert on his own sayso a sentence  $\phi$  in  $T$ 's model of  $M$  on the expectation that  $T$  with it in will do better than  $T$  with it not in, and subject in turn to its removal in the face of a subsequently disappointing performance by  $T$ . This is a point to make something of. Here is what we make of it:

- The extent to which a stipulation is held to the sufferance condition, the more it resembles a *working hypothesis*.
- The more a sentence operates as a working hypothesis, the more its introduction into a scientific theory is conditioned by *abductive considerations*.

Accordingly, despite its free standing in  $M$ , a stipulationist's  $\phi$  in  $T$  is bound by, as we may now say, *book-end* conditions, that is to say, conditions on *admittance* into  $T$  in the first place, and conditions on its *staying* in  $T$  thereafter. The conditions on going in are broadly abductive in character. The conditions on *staying in* are broadly – sometimes very broadly – confirmational in character. Since there is nothing remotely

abductive or confirmational in virtue of which a sentence is an  $\mathcal{F}$ -truth [fictive truth] on its author's sayso, radical pluralism must be our verdict here [Woods and Rosales, 2010a, pp. 375-376].

In conclusion, after having proposed a distinction between predicates that are load-bearing in a theory and those that are not, Woods and Rosales maintain that a predicate that is not load-bearing in a theory is a *façon de parler*: “For example, everyone will agree that the predicate ‘is a set’ is load-bearing in the mathematical theory of sets and that ‘is an abstract object’, if it occurs there at all, is a *façon de parler*. ‘Is an abstract object’ may well be load-bearing in the philosophy of mathematics, but no work-a-day mathematician need trouble with it. It generates no new theorems for him. Similarly, ‘reduces to logic’ is not load-bearing in number theory, notwithstanding the conviction among logicians that it is load-bearing in mathematical epistemology” [Woods and Rosales, 2010a, pp. 377–378]. Unfortunately the predicate “is a fiction” is non-load-bearing, or at best a *façon de parler*, in any scientific theory. At this point the conclusion is obvious, and I agree with it, since there is no concept of scientific fiction, the question of whether it is assimilable to or in some other way unifiable with the concept of literary fiction does not arise.

Elsewhere [Magnani, 2009, chapter three] I called the external scientific models “mimetic”,<sup>26</sup> not in a military sense, as camouflaged tools to trick the hostile eco-human systems, but just as structures that mimic the target systems for epistemic aims. In this perspective I described the centrality of the so called “disembodiment of the mind” in the case of semiotic cognitive processes occurring in science. Disembodiment of the mind refers to the cognitive interplay between internal and external representations, *mimetic* and, possibly, *creative*, where the problem of the continuous interaction between on-line and off-line (for example in inner rehearsal) intelligence can properly be addressed. In the subsection 3.4 below, we will see that this distinction parallels the one illustrated by Morrison between models which idealize (mirroring the target systems) and abstract models (more creative and finalized to establish new scientific intelligibility).

As I am trying to demonstrate in this whole section with the description of the above models based on common coding, I consider this interplay critical in analyzing the relation between meaningful semiotic internal resources and devices and their dynamical interactions with the externalized semiotic materiality already stored in the environment (scientific artifactual models, in this case). This external materiality plays a specific role in the interplay due to the fact that it exhibits (and operates through) its own cognitive constraints. Hence, minds are “extended” and artificial in themselves. It is at the level of that continuous interaction between on-line and off-line intelligence that I underlined the importance of what I called *manipulative abduction*.

Manipulative abduction, which is widespread in scientific reasoning [Magnani, 2009, chapter one] is a process in which a hypothesis is formed and evaluated resorting to a basically extra-theoretical and extra-sentential behavior that aims at cre-

<sup>26</sup> On the related problem of resemblance (similarity, isomorphism, homomorphism, etc.) in scientific modeling see below subsection 3.5.

ating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices. Manipulative abduction represents a kind of redistribution of the epistemic and cognitive effort to manage objects and information that cannot be immediately represented or found internally. An example of manipulative abduction is exactly the case of the human use of the construction of external models in the neural engineering laboratory I have outlined in the previous subsections, useful to make observations and “experiments” to transform one cognitive state into another to discover new properties of the target systems. Manipulative abduction also refers to those more unplanned and unconscious action-based cognitive processes I have characterized as forms of “thinking through doing” (cf. footnote 10 above).

### ***3.3 Are the In-Vitro Model or a Geometrical Diagram Fictions? Dynamic vs. Static View of Scientific Models***

In subsection 3.1 I have contended that Peirce, speaking of the model-based aspects of deductive reasoning, hypothesized there is an “experimenting upon this image [the external model/diagram] in the imagination”, so showing how human geometrical imagination is always triggered by a kind of prosthesis, the external model as an “external imagination”. Analogously, taking advantage of a fictional view on models and of the pretence theory Frigg [Frigg, 2010c, p. 266 ff.] interestingly sees imagination as an authorized intersubjective game of make-believe sanctioned by the “prop” (an object, for example material models, movies, paintings, plays, etc.) and its rules of generation. This theory also works as a metaphor of abductive processes, in terms of some concepts taken from the theory of literary and artistic fictions. Again, I think that it is neither necessary to adopt a fictionalist view in the case of science, nor the pretence theory adds something relevant to the issue. In the example I am illustrating in this section scientists in the lab do not pretend anything and are not engaged in the relative make-believe process, if not in the trivial sense that almost every human intersubjective interplay can be seen as such. The in-vitro networks of cultured neurons of our case or the Peircean Euclidean diagram used by the ancient Greek geometers are just the opposite of a mere fiction or of a generic make-believe interplay, they are instead more or less mimetic (possibly creative) external models which are expected to provide reliable information about the target system. They aim at discovering some new representations about the neurons in the first case and about the pure concepts of geometry in the second.

The reason of my skepticism can be illustrated taking advantage of some theses derived from classical Kantian philosophy and Thom’s mathematical semiphysics. Immanuel Kant was clearly aware of the interplay between internal and external models, exemplified in the case of a formal science like mathematics. In its transcendental terms, 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]. Hence,

for Kant models in science (in this case, of geometry) are first of all *constructions* that go beyond what the researcher simply “thinks”. We have seen that manipulative abduction is a kind of, usually model-based, abduction that exploits external models endowed with delegated (and often implicit) cognitive roles and attributes: 1. The model is external and the strategy that organizes the manipulations is unknown a priori. 2. The result achieved is new (if we, for instance in this geometrical case, 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]).<sup>27</sup>

*Iconicity* is central for Peirce, who analogously to Kant, maintains that “[...] philosophical reasoning is reasoning with words; while theorematic reasoning, or mathematical reasoning is reasoning with specially constructed schemata” [Peirce, 1931-1958, 4.233]; moreover, he uses diagrammatic and schematic as synonyms, thus relating his considerations to the Kantian tradition where schemata mediate between intellect and phenomena.<sup>28</sup> The following is the famous related passage in the *Critique of Pure Reason* (“Transcendental Doctrine of Method”):

Suppose a philosopher be given the concept of a triangle and he be left to find out, in his own way, what relation the sum of its angles bears to a right angle. He has nothing but the concept of a figure enclosed by three straight lines, and possessing three angles. However long he meditates on this concept, he will never produce anything new. He can analyse and clarify the concept of a straight line or of an angle or of the number three, but he can never arrive at any properties not already contained in these concepts. Now let the geometrician take up these questions. He at once begins by constructing a triangle. Since he knows that the sum of two right angles is exactly equal to the sum of all the adjacent angles which can be constructed from a single point on a straight line, he prolongs one side of his triangle and obtains two adjacent angles, which together are equal to two right angles. He then divides the external angle by drawing a line parallel to the opposite side of the triangle, and observes that he has thus obtained an external adjacent angle which is equal to an internal angle – and so on. In this fashion, through a chain of inferences guided throughout by intuition, he arrives at a fully evident and universally valid solution of the problem [Kant, 1929, A716-B744, pp. 578-579].

Here “intuition” is the Kantian word that expresses our present reference to what we call “external model”.

We can depict the situation of the philosopher described by Kant at the beginning of the previous passage taking advantage of some ideas coming from the catastrophe theory. As a human being who is not able to produce anything new relating to

<sup>27</sup> Of course in the case we are using diagrams to demonstrate already known theorems (for instance in didactic settings), the strategy of manipulations is often already available and the result is not new.

<sup>28</sup> Schematism, a fruit of the imagination is, according to Kant, “[...] an art concealed in the depths of the human soul, whose real modes of activity nature is hardly likely ever to allow us to discover, and to have open to our gaze” [Kant, 1929, A141-B181, p. 183]. Now we have at our disposal, thanks to epistemology and cognitive science, a lot of knowledge about the cognitive processes which correspond to Kantian schematism. On models as epistemic mediators in mathematics cf. [Boumans, 2012].

the angles of the triangle, the philosopher experiences a feeling of frustration (just like the Köhler's monkey which cannot keep the banana out of reach). The bad affective experience “deforms” the organism's regulatory structure by complicating it and the cognitive process stops altogether. The geometer instead “at once constructs the triangle” [the scientist constructs the model] that is, he makes an external representation of a triangle and acts on it with suitable manipulations. Thom thinks that this action is triggered by a “sleeping phase” generated by possible previous frustrations which then change the cognitive status of the geometer's available and correct internal idea of triangle (like the philosopher, he “has nothing but the concept of a figure enclosed by three straight lines, and possessing three angles”, but his action is triggered by a sleeping phase). Here the idea of the triangle is no longer the occasion for “meditation”, “analysis” and “clarification” of the “concepts” at play, like in the case of the “philosopher”. Here the inner concept of triangle – symbolized as insufficient – is amplified and transformed thanks to the sleeping phase (a kind of Kantian imagination active through schematization) in a prosthetic triangle to be put outside, in some external support. The instrument (here an external diagram) becomes the extension of an organ:

What is strictly speaking the end [...] [in our case, to find the sum of the internal angles of a triangle] must be set aside in order to concentrate on the means of getting there. Thus the problem arises, a sort of vague notion altogether suggested by the state of privation. [...] As a science, heuristics does not exist. There is only one possible explanation: the affective trauma of privation leads to a folding of the regulation figure. But if it is to be stabilized, there must be some exterior form to hold on to. So this anchorage problem remains whole and the above considerations provide no answer as to why the folding is stabilized in certain animals or certain human beings whilst in others (the majority of cases, needless to say!) it fails [Thom, 1988, pp. 63–64].<sup>29</sup>

### 3.4 *Confounding Static and Dynamic Aspects of the Scientific Enterprise*

Taking advantage of Thom's considerations, we can clearly see that the constructed external scientific model in the case of creative processes is exactly the opposite both of a fiction and of a generic process of make-believe (neither is a mere surrogate [Contessa, 2007] or a bare credible world [Sugden, 2000; Sugden, 2009]). It is instead a *regulatory tool stabilized* in “some exterior form”, a kind of a reliable anchorage, not intentionally established as fiction, like a romance writer could intentionally do, assessing the character of Harry Potter. In the epistemological fictionalism about models the use of the label “fiction” is usually legitimated by the fact that there are no empirical systems corresponding for example to the ideal pendulum (and its equation).

<sup>29</sup> A full analysis of the Köhler's chimpanzee getting hold of a stick to knock a banana hanging out of reach in terms of the mathematical models of the perception and the capture catastrophes is given in [Thom, 1988, pp. 62–64]. On the role of emotions, for example frustration, in scientific discovery cf. [Thagard, 2002].

Unfortunately the label sets up a paradox we can clearly see taking advantage of the case of scientific models seen as “missing systems”, another new metaphor that echoes the fictional one – indeed the description of a missing system might be a fiction. Thomson-Jones [Thomson-Jones, 2010] emphasizes that science is full of “descriptions of missing systems”, that at the end are thought as *abstract models*.<sup>30</sup> Further, Mäki [Mäki, 2009] usefully acknowledges that scientific models are “pragmatically and ontologically constrained representations”, and further complicates the missing systems framework adding a supplementary metaphoric conceptual apparatus: missing systems are also “surrogate” systems expressed as credible worlds, as models. Similar argumentations are advanced by Godfrey-Smith [Godfrey-Smith, 2009, pp. 114]: “To say that talk of model systems is a psychologically exotic way of investigating conditionals (and the like) is not itself to solve the problem. It is natural to think that the useable output we get from modeling is generally a conditional - a claim that if such-and such a configuration existed, it would behave in a certain way. The configurations in question, however, are usually known *not* to exist, so the problem of explaining the empirical usefulness of this kind of knowledge reappears”.

I contend that, at least in a discovery cognitive process, the missing system (Thomson-Jones) is not, paradoxically, the one represented by the “model”, but instead the target system itself, still more or less largely unknown and un-schematized, which will instead appear as “known” in a new way only after the acceptance of the research process results, thus admitted into the theory *T* and considered worth to *staying* in *T* thereafter.<sup>31</sup> The same can be said of models as configurations (Godfrey-Smith), which certainly are conditional, but at the same time not “known *not* to exist”, in Godfrey-Smith’s sense, because simply in the moment in which a scientific model is introduced in a discovery process it is instead exactly the only object we plausibly *know* to exist (for example a diagram in a blackboard, or a *in-vitro* artifact, or a mental imagery). Only in the framework of a strong metaphysical realism we can state that, once a final scientific result has been achieved, together with the description of the related experimental side, everything that does not fit that final structure is a fiction, and so models that helped reach that result itself. Morrison is pretty clear about the excessive habit of labeling fictional scientific models simply because they are superficially seen as “unrealistic”: “Although there is a temptation to categorize any type of unrealistic representation as a ‘fiction’, I have argued that this would be a mistake, primarily because this way of categorizing the use of unrealistic representations tells us very little about the role those representations play in producing knowledge” [Morrison, 2009, p. 133].

In the framework of an account of scientific representation in terms of partial structures and partial morphisms Bueno and French [Bueno and French, 2011, p. 27] admit that they agree in fact that an important role for models in science is to allow scientists to perform the so-called “surrogate” reasoning, but they add the

<sup>30</sup> Cartwright [Cartwright, 1983], more classically and simply, speaks of “prepared description” of the system in order to make it amenable to mathematical treatment.

<sup>31</sup> Cf. the previous subsection, on the problem of scientific model stipulation as subject to a *sufferance* constraint.

following constraint: “Indeed, we would claim that representing the ‘surrogate’ nature of this reasoning effectively rides on the back of the relevant partial isomorphisms, since it is through these that we can straightforwardly capture the kinds of idealizations, abstractions, and inconsistencies that we find in scientific models”. So to say, we can speak of surrogates, fictions, credible worlds, etc., but it is only through the suitable partial isomorphism we can detect after a success of the model, that we can be assured to be in presence of a scientific representation or model.

Further, Kuorikoski and Lehtinen [Kuorikoski and Lehtinen, 2009, p. 121] contend that: “The epistemic problem in modelling arises from the fact that models always include false assumptions, and because of this, even though the derivation within the model is usually deductively valid, we do not know whether our model-based inferences reliably lead to true conclusions”. However, the false premises (also due to the presence in models of both substantive and auxiliary assumptions) are not exploited in the cognitive process, because, in various heuristic processes, only the *co-exact* ones are exploited.<sup>32</sup> Moreover, some false assumptions are considered as such *only if* seen in the light of the still “to be known” target system, and so they appear false only in a *post hoc* analysis, but they are perfectly true in the model itself in its relative autonomy during the smart heuristic cognitive process related to its exploitation. So various aspects of the model are the legitimately true basis for the subsequent exploration of its behavior and performance of the abductions to plausible hypotheses concerning the target system. I agree with Morrison: “I see this not as a logical problem of deriving true conclusions from false premises but rather an epistemic one that deals with the way false representations transmit information about concrete cases” [Morrison, 2009, p. 111].<sup>33</sup>

<sup>32</sup> The notion of co-exact proprieties, introduced by Manders [Manders, 2008], is worth to be further studied in fields that go beyond the realm of discovery processes of classical geometry, in which it has been nicely underscored. Mumma [Mumma, 2010, p. 264] illustrates that Euclid’s diagrams contribute to proofs only through their co-exact properties. Indeed “Euclid never infers an exact property from a diagram unless it follows directly from a co-exact property. Exact relations between magnitudes which are not exhibited as a containment are either assumed from the outset or are proved via a chain of inferences in the text. It is not difficult to hypothesize why Euclid would have restricted himself in such a way. Any proof, diagrammatic or otherwise, ought to be reproducible. Generating the symbols which comprise it ought to be straightforward and unproblematic. Yet there seems to be room for doubt whether one has succeeded in constructing a diagram according to its exact specifications perfectly. The compass may have slipped slightly, or the ruler may have taken a tiny nudge. In constraining himself to the co-exact properties of diagrams, Euclid is constraining himself to those properties stable under such perturbations”.

<sup>33</sup> Further information about the problem of the mapping between models and target systems through *interpretation* are provided by Contessa [Contessa, 2007, p. 65] – interpretation is seen as more fundamental than surrogate-reasoning: “The model can be used as a generator of hypotheses about the system, hypotheses whose truth or falsity needs to be empirically investigated”. By using the concept of interpretation (analytically and not hermeneutically defined) the author in my opinion also quickly adumbrates the creative aspects in science, that coincide with the fundamental problem of model-based and manipulative abduction (cf. [Magnani, 2009, chapters one and two]).



In sum, I think it is misleading to analyze models in science by adopting a confounding mixture of static and dynamic aspects of the scientific enterprise. Scientific models in a static perspective (for example when inserted in a textbook) certainly appear – but just appear – fictional, because they are immediately compared with the target systems and their complicated experimental apparatuses: in this case also the *ideal* character of models becomes manifest and so the *explanatory* function of them (cf. [Weisberg, 2007]). Contrarily, scientific models seen inside the living dynamics of scientific creativity, which is the key topic of epistemology at least since Karl Popper and Thomas Kuhn, appear *explicit* and *reproducible* machineries intentionally built and manipulated to the gnoseological aims of increasing scientific knowledge *not yet available*.

Morrison [Morrison, 2009] is certainly not inclined to see models as fictions because she emphasizes that in science they are specifically related to (“finer graded”) ways of understanding and explaining “real systems”, far beyond their more collateral predictive capabilities and their virtues in approximating. She indeed further clarifies that the models which is appropriate to label as *abstract* resist – in the so-called process of de-idealization – corrections or relaxing of the unrealistic assumptions (such as in the case of mathematical abstractions or when models furnish the sudden chance for the applicability of equations), because they are “necessary” to arrive to certain results. The fact that in these models “relevant features” are subtracted to focus on a single – and so isolated – set of properties or laws, as stressed by Cartwright [Cartwright, 1989], is not their central quality, because what is at stake is their capacity to furnish an overall new depiction of an empirical (and/or theoretical, like in case of mathematics or logic) framework: “[...] We have a description of a physically unrealizable situation that is required to explain a physically realizable one” (p. 130).

Other models, easier to define, which is better to classify as *idealizations*, allow “[...] for the addition of correction factors that bring the model system closer (in representational terms) to the physical system being modelled or described” [Morrison, 2009, p. 111]. It is for example the case of simple pendulum, where we know how to add corrections to deal with concrete phenomena. Idealizations distort or omit properties, instead abstractions introduce a specific kind of representation “that is not amenable to correction and is necessary for explanation/prediction of the target system” (p. 112), and which provides information and transfer of knowledge. Morrison’s characterization of scientific models as abstract is in tune with my emphasis on models as *constitutive*, beyond the mere role played by models as idealizations, which instead allow corrections and refinements (cf. below, subsection 3.6). In this perspective, “abstract” models, either related to prepare and favor mathematization or directly involving mathematical tools, have to be intended as poetic ways of producing new intelligibility of the essential features of the target systems phenomena, and not mere expedients for facilitating calculations. If idealization *resembles* the phenomena to be better understood, abstract models *constitute* the resemblance itself, as I will illustrate in the following subsection.

When Mäki [Mäki, 2009, p. 31] contends that “It may appear that a fantastically unreal feature is added to the model world, but again, what happens is that one

thereby removes a real-world feature from the model world, namely the process of adjustment”, I have to note that, at least in various creative processes, the model is not necessarily implemented through “removal” or “neutralization” of real-world features, because some features of the target system – that is the supposed to be real world – have simply not been discovered yet, and so, paradoxically, they are the ones still “missing”. Consequently it is impossible to imagine that some aspects of the model derive from a removal of features of the real world, that can just be those features that will derive later on exactly thanks to that cognitive process that constructed the model itself to reach that objective. At the same time, and for the same reason, it is difficult to always state that models depict a “surrogate” systems, because the systems we want to subrogate *are largely not yet known*.

### 3.5 *Resemblance and Feyerabend’s Counterinduction*

Even the concept of resemblance (similarity, isomorphism, homomorphism, etc.) as it is employed in the epistemological framework of missing systems (and related topics, fictions, surrogate systems, credible world, make-believe models, etc.) is in part misleading. “*M* resembles, or corresponds to, the target system *R* in suitable respects and sufficient degrees. This second aspect of representation enables models to serve a useful purpose as representatives: by examining them as surrogate systems one can learn about the systems they represent” [Mäki, 2009, p. 32]: I contend that resemblance is constitutively partial *also* because it is basically impossible to appropriately resemble things that are not yet known.<sup>34</sup>

It is not always acknowledged in the current literature that isomorphism, homomorphism and similarity with the target systems *are not* necessarily established – so to say – a priori, because the target system has still to be built. Actually – this is an important point – it is just the work of models that of creating, in a poetic way, the “resemblance” to the target system. Some discovered features of the target system resemble the model not because the model resembled them a priori but only *post hoc*, once discovered thanks to the modeling activity itself, in so far as resemblance has been *instituted* by the model: the new features appear well-defined only in the static analysis of the final developed theory. It is at this stage that resemblance acquires the actual status of resemblance, in the common sense of the word: similarity of two *given* entities/structures. Morrison too contends that “To say that fictional models are important sources of knowledge in virtue of a particular kind of similarity that they bear to concrete cases or systems is to say virtually nothing about how they do that. Instead what is required is a careful analysis of the model itself to uncover the kind of information it yields and the ways in which that information can be used to develop physical hypotheses” [Morrison, 2009, p. 123].

---

<sup>34</sup> On the puzzling relationships between similarity and representations, in the framework of intentionality, cf. [Giere, 2007].

In this perspective we paradoxically face the opposite of the received view, it is the newly known target system that resembles the model, which itself originated that resemblance.<sup>35</sup> Often models are useful to discover new knowledge just because they do not – or scarcely – resemble the target system to be studied, and are instead built to the aim of finding a new general capacity to make “the world intelligible”.<sup>36</sup>

In *Against Method* [Feyerabend, 1975], Feyerabend attributes a great importance to the role of contradiction, against the role of similarity. He establishes a “counterrule” which is the opposite of the neopositivistic one that it is “experience” (or “experimental results”) which measures the success of our theories, a rule that constitutes an important part of all theories of corroboration and confirmation. The counterrule “[...] advises us to introduce and elaborate hypotheses which are inconsistent with well-established theories and/or well-established facts. It advises us to proceed counterinductively” [Feyerabend, 1975, p. 20]. Counterinduction is seen more reasonable than induction, because appropriate to the needs of creative reasoning in science: “[...] we need a dream-world in order to discover the features of the real world we think we inhabit” (p. 29). We know that counterinduction, that is the act of introducing, inventing, and generating new inconsistencies and anomalies, together with new points of view incommensurable with the old ones, is congruous with the aim of inventing “alternatives” (Feyerabend contends that “proliferation of theories is beneficial for science”), and very important in all kinds of creative reasoning. Feyerabend stresses the role of “dreaming”, but these dreams are Galileo’s dreams, they are not fictions: as I have already pointed out Feyerabend clearly distinguished between scientific dreams (as modeling) and propaganda, that can instead be organized thanks to fictions, inconsistent thought experiments, mistakes, aggressive fallacies, and so on, but that do not play any epistemic role in the restricted cognitive process of scientific discovery, I have called “epistemic” warfare.<sup>37</sup>

Coming back to the problem of models as surrogates, Mäki [Mäki, 2009, p. 35] says:

The model functions as a surrogate system: it is construed and examined with a desire to learn about the secrets of the real world. One yearns for such learning and sets out to build a model in an attempt to satisfy the desire. Surrogate models are intended, or can be employed to serve, as bridges to the world.

---

<sup>35</sup> I endorse many of the considerations by Chakravartty [Chakravartty, 2010], who stresses the unwelcome division between informational and functional perspective on models and representations in science, which negatively affects the epistemology of scientific modeling.

<sup>36</sup> I am convinced that knowledge about concepts such as resemblance, imaginability, conceivability, plausibility, persuasiveness, credit worthiness [Mäki, 2009, pp. 39–40] would take advantage of being studied in the framework of the rigorous and interdisciplinary field of abductive cognition [Magnani, 2009], which surprisingly is largely disregarded in the studies of the “friends of fiction”, with the exception of Sugden [Sugden, 2000; Sugden, 2009].

<sup>37</sup> On Galileo’s mental imagery, cf. below, subsection 3.6.

First, I would add some auxiliary notes to the expression “secrets of the real world”. I would warn about the preferability of being post-Kantian instead than pre-Kantian by admitting that, through science, we are *constructing* our rational knowledge of the world, which consequently is still an objective world independent of us, but constructed. If we say we build surrogate systems to learn about the secret of nature, a strong realist assumption seems to be presupposed: the models would be surrogates because they are not “reliably reflecting the true reality of the world we are discovering”. We rejoin Giere’s observation I already quoted above (section 2) who suspects fictionalists are paradoxically obsessed by “the truth, the whole truth, and nothing but the truth”: scientific theories would reflect this hyper-truth that in turn would reflect true reality (curious! Is not science the realm or self-correcting truths?)<sup>38</sup> In this way it becomes easy to say that everything else in science different from complete established true theories – which would reflect “real world” – is fiction, surrogate, belief, mere credible world, etc.

I would reserve the label of surrogate models to those models employed in some “sciences” that fail in providing satisfactory knowledge about target systems. “There is a long tradition in economics of blaming economists for failing in just this way: giving all their attention to the properties of models and paying none to the relations of the model worlds to the real world” [Mäki, 2009, p. 36]. Mäki calls the systems described by such models “substitute systems”: I will just reserve for them the expression “surrogate systems”, because they fake a scientific knowledge that is not satisfactorily achieved, from various perspectives.

I argued above about the epistemological poverty of the concept of model as make-believe: indeed I have already said that make believe processes trivially occur in almost every human intersubjective interplay. Here I can further stress that the idea of credible world is very wide: every cognitive process that aims at providing scientific – but also non scientific – knowledge aims at the same time at providing credible worlds. The problem in science is how to construct the subclass of *epistemologically* credible worlds, that is, *scientific* models, which successfully lead to scientific theories. In this spirit Sugden [Sugden, 2009, p. 10] usefully suggests that an epistemologically “good” credible world would have to be provided by models that are able to trigger hypotheses about the “causation of actual events”, that is in cases in which “the fictional world of the model is one that *could* be real”. Cartwright’s classical model [Cartwright, 2009a] concerning capacities is fruitfully adopted:

For her, the function of a model is to *demonstrate the reality* of a capacity by isolating it – just as Galileo’s experiment demonstrates the constancy of the vertical component of the acceleration of a body acted on by gravity. Notice how Cartwright speaks of *showing that C* has the capacity to produce *E*, and of deriving this conclusion from *accepted principles*. A satisfactory isolation, then, allows a real relationship of cause and effect to be demonstrated in an environment in which this relationship is stable. In more natural conditions, this relationship is only a latent capacity which may be

<sup>38</sup> We should not forget what Morrison reminds us: “Laws are constantly being revised and rejected; consequently, we can never claim that they are true or false” [Morrison, 2009, p. 128].

switched on or off by other factors; but the capacity itself is stable across a range of possible circumstances. Thus, the model provides a “theoretical grounding” for a general hypothesis about the world [Sugden, 2009, p. 20].

Sugden prudently considers too strong these perspective on models as tools for *isolating* the “capacities” of causal factors in the real world, and provides other conceptual devices to save various aspects of epistemological – supposed to be weak – “sciences”, for example some parts of biology, psychology, or economics, which not ever fulfill the target of revealing capacities. To save these sciences he says that models can simply provide “conceptual explorations”, which ultimately contribute to the development of genuinely explanatory theories or credible counterfactual worlds which can trigger inductive (or “abductive”) inferences to explain the target systems. I think that it is virtuous to be prudent about strong methodological claims such as the ones advanced by Cartwright, but the epistemological problem remains open: in the cases of models as conceptual exploration are they used to depict credible worlds able to reach satisfactory theorization of target systems, or are they just providing ambitious but unjustified hypotheses, devoid of various good epistemological requisites?

Adopting Cartwright’s rigid demarcation criterium clearly and recently restated in “If no capacities then no credible worlds” [Cartwright, 2009a], it would seem that no more citizenship is allowed to some post-modern exaggeration in attributing the label “scientific” to various proliferating areas of academic production of knowledge, from (parts of) psychology to (parts of) economics, and so on, areas which do not – or scarcely – accomplish the most common received epistemological standards, for example, the *predictivity* of the phenomena that pertain the explained systems. Are we sure that this demarcation is too rigid or it is time to criticize some excess in the proliferation of models supposed to be “scientific”? It is in this perspective that the epistemological use of the so-called credible worlds appears theoretically suspect, but ideologically clear, if seen in the “military” framework of the academic struggle between disciplines, dominated – at least in my opinion – by a patent proliferation of “scientific” activities that just produce bare “credible” or “surrogate” models, looking aggressively for scientificity, when they actually are, at the best, fragments of *bad philosophy*.<sup>39</sup>

<sup>39</sup> An example is furnished by the precarious condition of various parts of psychological research. Miller [Miller, 2010, p. 716] explores three contentions: “[...] that the dominant discourse in modern cognitive, affective, and clinical neuroscience assumes that we know how psychology/biology causation works when we do not; that there are serious intellectual, clinical, and policy costs to pretending we do know; and that crucial scientific and clinical progress will be stymied as long as we frame psychology, biology, and their relationship in currently dominant ways”. He further rigorously illustrates the misguided attempts to localize psychological function via neuroimaging and the misunderstandings about the role of genetics in psychopathology, sadly intertwined with untoward constraints on health-care policy and clinical service delivery.

### 3.6 *Galileo's Modeling Vindicated*

Weisberg [Weisberg, 2007, p. 642]<sup>40</sup> maintains that “Galilean idealization is the practice of introducing distortions into theories with the goal of simplifying theories in order to make them computationally tractable. One starts with some idea of what a non-idealized theory would look like. Then one mentally and mathematically creates a simplified model of the target”. I would like to advance a suspect about this canonical treatment of Galileo, and provide some reasons that explain my perplexity.

When Galileo illustrates an imaginary model concerning the problem of falling bodies, he provides a kind of smart mental modeling. Let us religiously follow the text of the creator of modern science on this subject:

SALV. But, even without further experiment, it is possible to prove clearly, by means of a short and conclusive argument, that a heavier body does not move more rapidly than a lighter one provided both bodies are of the same material and in short such as those mentioned by Aristotle. But tell me, Simplicio, whether you admit that each falling body acquires a definite speed fixed by nature, a velocity which cannot be increased or diminished except by the use of force [violenza] or resistance.

SIMP. There can be no doubt but that one and the same body moving in a single medium has a fixed velocity which is determined by nature and which cannot be increased except by the addition of momentum [impeto] or diminished except by some resistance which retards it.

SALV. If then we take two bodies whose natural speeds are different, it is clear that on uniting the two, the more rapid one will be partly retarded by the slower, and the slower will be somewhat hastened by the swifter. Do you not agree with me in this opinion?

SIMP. You are unquestionably right.

SALV. But if this is true, and if a large stone moves with a speed of, say, eight while a smaller moves with a speed of four, then when they are united, the system will move with a speed less than eight; but the two stones when tied together make a stone larger than that which before moved with a speed of eight. Hence the heavier body moves with less speed than the lighter; an effect which is contrary to your supposition. Thus you see how, from your assumption that the heavier body moves more rapidly than the lighter one, I infer that the heavier body moves more slowly.

SIMP. I am all at sea because it appears to me that the smaller stone when added to the larger increases its weight and by adding weight I do not see how it can fail to increase its speed or, at least, not to diminish it.

SALV. Here again you are in error, Simplicio, because it is not true that the smaller stone adds weight to the larger.

SIMP. This is, indeed, quite beyond my comprehension. [Galilei, 1914, pp. 62–63].

Gendler nicely summarizes this kind of Galilean mental modeling stressing that we are dealing with an admirable example of *Gedankenexperiment*

<sup>40</sup> Weisberg distinguished between various kinds of idealization: Galilean, minimalist (still devoted to reveal the most important causal powers at stake), and multiple-models (devoted of a single representation ideal, widespread for example in biology and social science).

(thought experiment) in which we imagine that a heavy and a light body are strapped together and dropped from a significant height:

What would the Aristotelian expect to be the natural speed of their combination? On the one hand, the lighter body should slow down the heavier one while the heavier body speeds up the lighter one, so their combination should fall with a speed that lies between the natural speeds of its components. (That is, if the heavy body falls at a rate of 8, and the light body at a rate of 4, then their combination should fall at a rate between the two [...].) On the other hand, since the weight of the two bodies combined is greater than the weight of the heavy body alone, their combination should fall with a natural speed greater than that of the heavy body. (That is, if the heavy body falls at a rate of 8 and the light body with a rate of 4, their combination should fall at a rate greater than 8.) But then the combined body is predicted to fall both more quickly, and more slowly, than the heavy body alone. The way out of this paradox is to assume that the natural speed with which a body falls is independent of its weight: “both great and small bodies [...] are moved with like speeds” [Gendler, 1998, p. 403].

Is this modeling a fiction, a surrogate, an idealization, an abstraction, a credible world of the target system? Surely these attributes do not appropriately characterize this Galileo’s epistemic act, which cognitively attacks the Aristotelian views on motion. Let us explain why. For the Aristotelian, the daily experience seems to confirm that heavier bodies fall faster than the lighter ones. Nevertheless, when the Aristotelian sees two stones of different weights fall the ground with similar speeds, this requires an explanation.<sup>41</sup> Two auxiliary assumptions can be provided, the Galilean one in terms of air resistance, the Aristotelian one which complains that the bodies have not been dropped from a height sufficiently great. What the Galilean thought experiment provides to the Aristotelian is not a new empirical knowledge of the external world but a sudden new belief, or a “conceptual reconfiguration”, concerning the independency between speed and weight of falling bodies, and the *kind of thing* natural speed might be as a new *physical property*, like Gendler says [Gendler, 1998, pp. 408–409].

Given the fact the modeling activity provided by this thought experiment is not posterior to the conceptual “reconfiguration” of the empirical data, it could hardly be classified as fictional or as a surrogate of them, it is instead *constitutive* of the possible reconfiguration itself: “Prior to contemplation of the case, there was no room on the Aristotelian picture for the thought that natural speed might be constant, not varying – that it might be dependent not on some specific features of the body in question, but only on the fact that it is a body at all” [Gendler, 1998, p. 412]. The old Aristotelian idea of natural speed does not make sense anymore “like phlogiston, it disappears into the ether of abandoned concepts” (cit.) The model provided by the thought experiment is not a simple way of modifying the Aristotelian perception of falling bodies, but a transformation of the “schematization” of the percepts themselves, to use the Kantian efficacious word, which makes them intelligible in a novel way. And, like experiments in science, this good thought experiment is not

<sup>41</sup> Philosophy of science has often stressed that theories are undetermined by evidence, like for example the conventionalist tradition teaches us [Magnani, 2001b, chapter five].

evanescent and fuzzy, but clear, *repeatable*, and *sharable*, in so far as it can involve unambiguous constructive representations in various human agents.

In this case the model is “crucial”: “There will, no doubt, be many cases where the role of the imagery is simply heuristic. But there will also be cases where the role of the imagery is [...] epistemically crucial” [Gendler, 2004, p. 1161].<sup>42</sup> This “crucial” creative role is also stressed by Nersessian [Nersessian, 1993, p. 292] who, describing Mach’s seminal ideas on the *Gedankenexperiment*, reminds us that “[...] while thought experimenting is a truly creative part of scientific practice, the basic ability to construct and execute a thought experiment is not exceptional. The practice is highly refined extension of a common form of reasoning [...] by which we grasp alternatives, make predictions, and draw conclusions about potential real-world situations” [Nersessian, 1993, p. 292].

Instead of seeing Galilean model as a fiction, it has to be considered an *actual* representation,<sup>43</sup> which helps discover – and justify – in this case in a precise *model-based* non-propositional way, what sorts of motions (and objects) we think plausible in the world. The door that provides access to further mathematical refinement and experimental research concerning the target system is finally opened. It will be only after having fruitfully built the complete Galilean mathematized theory of motion that the mental model provided by the thought experiment in question can appear fictional, a surrogate, and so on. Moreover, it is only at this later stage that also a clear concept of approximation (and, in turn, of de-idealization) of related models will acquire a rigorous and complete sense.<sup>44</sup> No distortions are present in the pre-supposed “idealization” of this Galilean thought experiment, simply because, the new schematization of the target is the fruit itself of the modeling activity, and we cannot provide a distortion of objects/targets that are not yet available. If still we want to say that the model shows itself as an idealization, this is simply because it belongs to modern physics, which on the whole, Galileo teaches us, idealizes.

<sup>42</sup> The basic epistemological and cognitive aspects of thought experiments are nicely illustrated by Arcangeli [Arcangeli, 2010], who stresses their role in producing new knowledge and the useful distinction between their icaistic or recreative character.

<sup>43</sup> “[...] the person conducting the experiment asks herself: ‘What would I say/judge/expect were I to encounter circumstances XYZ?’ and then finds out the (apparent) answer. This technique is common in linguistics, where the methodology is used to ascertain the grammaticality of sentences, the meanings of phrases, the taxonomic categories of words, and so on. And it is, on one view at least, a central element of moral reasoning: we think about particular imaginary cases, observe the judgements that they evoke in us, and use these judgements as fixed points in developing our moral theories” [Gendler, 1998, p. 414].

<sup>44</sup> A deep analysis of the relationships between idealization, approximation (and de-idealization), which is also in part in tune with my observations above, is provided by Portides [Portides, 2007, p. 708]: “I employ this analysis of the process of construction of representational models to demonstrate that idealisation, and its converse process of de-idealisation, is present at every level of scientific theorising whereas the concept of approximation becomes methodologically valuable, and epistemically significant, either when a tractable mathematical description of a de-idealising factor is needed or after a certain point in the process is reached when a given theoretical construct (i.e. a scientific model) may be proposed for the representation of a physical system”.



A further remark which takes advantage of Cartwright's epistemology of models and capacities can be useful to grasp the point about Galilean mental modeling. Treating Sugden's problem of models as credible worlds (that I have quoted above in subsection 3.4), Cartwright contends that "[...] the license to move from the results in the model about what happens when a cause is exercised without impediment to a contribution that the cause will make in all situations of some designated category depends on the assumption that the cause has a stable contribution to make, and that assumption must be supported by evidence from elsewhere. This is part of the way in which Sugden's own view relies on the logic of capacities" [Cartwright, 2009a, pp. 53–54]. It is very easy for Cartwright to add that capacities in science are characterized by some additional "premises": 1) "stable contribution" of the envisaged cause (eventually to be measured) in the real-world situation is not necessarily the same it does in the model, when we know that some other cause can have generated the effect in question; 2) the contribution the capacity makes in the model, the result "is exported to understand or predict in real-world situations where the cause that carries that capacity operates even when we do not expect the overall results to be the same in those situations that have results similar to those situations as they are in the models" (p. 54).

This is an important point, Cartwright says, because Sugden's account based on credible worlds simply looks at the real-world situation that presents results similar to those in the model and then infers by *abduction* that the causes are the same. Here a bad example of the fallacy of affirming the consequent is committed: we face the cognitive activity of inferring from the same effect to the same cause (pp. 54–55) and not, on the contrary, the fact that "whenever the same cause appears as in the model, the same effect will appear", because we can do this given the fact the model is based on a robust hypothesis about the complex relationships between cause and effect. Indeed, in this case, the abduction as "inferring from same effect to same cause" is highly uncertain, and it does not tell us that the model furnishes a stable contribution, which instead should only be related to the level of abstraction at which to describe the cause and the effect. It is the presence of more abstract concepts which describe the causes and effects that qualifies the epistemological quality of the model.

In the Galilean thought experiment I have illustrated above the bodies are envisaged as masses, and gravity is implied: this is exactly what is at the basis of the fertile exportation of conclusions from the model to the world, and of the possibility of finding a suitable schematization through mathematization. In the Galilean case, and by adopting a dynamic perspective on science, abduction is good and creative *because* we deal with the abductive process that concerns the *first* construction of modern physics. Otherwise, if we already possess the complete laws of Galilean physics – by adopting in this case a static perspective – a related model exports to the real situation thanks to a causal explanation through de-idealization. Indeed, Cartwright observes "Say we have a model about the planetary system. In the model we deduce that planets are caused by gravitational attraction to accelerate towards the sun. Is the motion of cannonballs towards the earth a similar effect so that we might do an abduction to similar causes? It is if we describe both the cannonballs

and the planets as compact masses. Otherwise the abduction is farfetched” (p. 57). In the perspective of this important distinction the epistemological divergence between static and dynamic aspects of science is still at stake.

In the case discussed above, of the model as a generic credible world, the model is instead “shallow” – as it happens in the case of simple analogue economic models – because it does not lead to discover proper capacities – in Cartwright’s sense – and unfortunately basic principles are neither available nor “foreseeable” through a working discovering modeling process, to which the model itself eventually strategically belongs. In the case of these shallow models Cartwright nicely concludes “the worry is not just that the assumptions are unrealistic; rather, they are unrealistic in just the wrong way” (p. 57). In this case models certainly are isolating devices, but they isolate in the wrong way, and induction – in Sugden’s sense, even if cautious – from the model to a real situation results to be a clear hasty generalization. This does not mean that these shallow models do not provide knowledge about target systems, but this knowledge is very limited and unsatisfactory in the light of the decent epistemological standards in terms of Cartwright’s capacities.

To conclude, coming back to the problem of fictionalism and its discontents, Galileo is explicitly clear about the distinction between science (he calls “philosophy” in the following celebrated passage) and literary fiction:

In Sarsi<sup>45</sup> I seem to discern the firm belief that in philosophizing one must support oneself upon the opinion of some celebrated author, as if our minds ought to remain completely sterile and barren unless wedded to the reasoning of some other person. Possibly he thinks that philosophy is a book of fiction by some writer, like the *Iliad* or *Orlando Furioso*, productions in which the least important thing is whether what is written there is true. Well, Sarsi, that is not how matters stand. Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth [Galilei, 1957, pp. 237–238].<sup>46</sup>

<sup>45</sup> Lothario Sarsi of Siguenza is the pseudonym of the Jesuit Orazio Grassi, author of *The Astronomical and Philosophical Balance*. In *The Assayer*, Galileo weighs the astronomical views of Orazio Grassi about the nature of the comets, and finds them wanting [Galilei, 1957, p. 231].

<sup>46</sup> As Bertolotti [Bertolotti, 2012] in this volume observes, the quotation obviously should not be used as an authority weapon against those who advocate the fictional nature of scientific models, because we would commit a fallacy, given the fact that to affirm that scientific models are fictions does not coincide with saying that the whole scientific endeavor has a fictional nature. Thus, the use of this quotation does not aim at getting definitively rid of fictionalism through the authority of one of the founding fathers of modern science.

## 4 Conclusion

In this paper I have contended that scientific models are not fictions. I have argued that also other various related epistemological approaches to model-based scientific cognition (in terms of surrogates, credible worlds, missing systems, make-believe) present severe inadequacies, which can be detected taking advantage of recent cognitive research in scientific labs and of the concept of manipulative abduction. In the meantime the illustrated critique, also performed in the light of distributed cognition, offered new insight on the analysis of the two main classical attributes given to scientific models: abstractness and ideality. A further way of delineating a more satisfactory analysis of fictionalism and its discontents has been constructed by proposing the concept of “epistemic warfare”, which sees scientific enterprise as a complicated struggle for rational knowledge in which it is crucial to distinguish epistemic (for example scientific models) from extra-epistemic (for example fictions, falsities, propaganda) weapons. I conclude that, in scientific settings, when models are fictions, it is because they were simply discarded, as heuristic failed steps, abandoned by applying a kind of negation as failure. I have also illustrated that it is misleading to analyze models in science by confounding static and dynamic aspects of the scientific enterprise: indeed the static perspective leads to an overemphasis of the possible fictional character of models because the creative/factive role of modeling is candidly or intentionally disregarded.

I have adopted some thoughts of two classical authors, which are of help in dealing with scientific modeling. Feyerabend’s useful concept of counterinduction in criticizing the role of resemblance in model-based cognition has been considered. In this perspective I have paradoxically reached the opposite of the received view: it is the newly known target system that resembles to the model, which itself originated that resemblance. Finally, to pleasantly try to give rid of fictionalism, the authoritative “voice” of Galileo is exploited: 1) the Galileo’s thought experiment I have illustrated shows how modeling in science (natural philosophy, for Galileo) is *constitutive* of central aspects of the target system that is studied, and surely it is not a fiction; 2) Galileo also explicitly says in *The Assayer* that we do not have absolutely to think that science “is a book of fiction by some writer, like the Iliad or Orlando Furioso, productions in which the least important thing is whether what is written there is true”.

**Acknowledgements.** For the instructive criticisms and precedent discussions and correspondence that helped me to develop my critique of fictionalism, I am indebted and grateful to Mauricio Suárez, Shahid Rahman, John Woods, Alirio Rosales, and to my collaborators Emanuele Bardone and Tommaso Bertolotti.

## References

- [Arcangeli, 2010] Arcangeli, M.: Imagination in Thought Experimentation: Sketching a Cognitive Approach to Thought Experiments. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 571–587. Springer, Heidelberg (2010)

- [Bardone, 2011] Bardone, E.: Seeking Chances. From Biased Rationality to Distributed Cognition. Springer, Heidelberg (2011)
- [Bardone, 2012] Bardone, E.: Not by Luck Alone: The Importance of Chance-Seeking and Silent Knowledge in Abductive Cognition. In: Magnani, L., Li, L. (eds.) *Philosophy and Cognitive Science*. SAPERE, vol. 2, pp. 187–205. Springer, Heidelberg (2012)
- [Barsalou, 2008a] Barsalou, L.W.: Cognitive and neural contributions to understanding the conceptual system. *Current Directions in Psychological Science* 17(2), 91–95 (2008)
- [Barsalou, 2008b] Barsalou, L.W.: Grounded cognition. *Annual Review of Psychology* 59, 617–645 (2008)
- [Bertolotti, 2012] Bertolotti, T.: From Mindless Modeling to Scientific Models. The Case of Emerging Models. In: Magnani, L., Li, L. (eds.) *Philosophy and Cognitive Science*. SAPERE, vol. 2, pp. 77–106. Springer, Heidelberg (2012)
- [Bokulich, 2011] Bokulich, A.: How scientific models can explain. *Synthese* 1, 33–45 (2011)
- [Boumans, 2012] Boumans, M.J.: Mathematics as quasi-matter to build models as instruments. In: Weber, M., Dieks, D., Gonzalez, W.J., Hartman, S., Stadler, F., Stöltzner, M. (eds.) *Probabilities, Laws, and Structures*, pp. 307–318. Springer, Heidelberg (2012)
- [Bueno and French, 2011] Bueno, O., French, S.: How theories represent. *The British Journal for the Philosophy of Science* (2011), doi:10.1093/bjps/axr010
- [Cartwright, 1983] Cartwright, N.: *How the Laws of Physics Lie*. Oxford University Press, Oxford (1983)
- [Cartwright, 1989] Cartwright, N.: *Nature's Capacities and Their Measurement*. Oxford University Press, Oxford (1989)
- [Cartwright, 2009a] Cartwright, N.: If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis* 70, 45–58 (2009)
- [Cartwright, 2009b] Cartwright, R.: Models: Parables v. fables. *Insights* 1(8), 2–10 (2009)
- [Chakravartty, 2010] Chakravartty, A.: Informational versus functional theories of scientific representation. *Synthese* 172, 197–213 (2010)
- [Chandrasekharan, 2009] Chandrasekharan, S.: Building to discover: a common coding model. *Cognitive Science* 33, 1059–1086 (2009)
- [Clark, 1978] Clark, K.L.: Negation as failure. In: Gallaire, H., Minker, J. (eds.) *Logic and Data Bases*, pp. 94–114. Plenum, New York (1978)
- [Contessa, 2007] Contessa, G.: Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science* 74, 48–68 (2007)
- [Contessa, 2010] Contessa, G.: Scientific models and fictional objects. *Synthese* 172, 215–229 (2010)
- [da Costa and French, 2003] da Costa, N.C., French, S.: *Science and Partial Truth. A Unitary Approach to Models and Scientific Reasoning*. Oxford University Press, Oxford (2003)
- [De Cruz and De Smedt, 2011] de Cruz, H., de Smedt, J.: Mathematical symbols as epistemic actions. *Synthese* (2011), doi:10.1007/s11229-010-9837-9
- [Feyerabend, 1975] Feyerabend, P.: *Against Method*. Verso, London-New York (1975)
- [Fine, 2009] Fine, A.: Fictionalism. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 19–36. Routledge, London (2009)
- [French, 2010] French, S.: Keeping quiet on the ontology of models. *Synthese* 172, 231–249 (2010)
- [Freud, 1953-1974] Freud, S.: *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press, London (1953); Translated by Strachey, J. in collaboration with Freud, A., et al.
- [Frigg, 2010a] Frigg, R.: Fiction and scientific representation. In: Frigg, R., Hunter, M.C. (eds.) *Beyond Mimesis and Nominalism: Representation in Art and Science*, pp. 97–138. Springer, Heidelberg (2010)

- [Frigg, 2010b] Frigg, R.: Fiction in science. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 247–287. Philosophia Verlag, Munich (2010)
- [Frigg, 2010c] Frigg, R.: Models and fiction. *Synthese* 172, 251–268 (2010)
- [Galilei, 1914] Galilei, G.: *Dialogues Concerning Two New Sciences* (1638). Mac Millan, New York (1914); Translated from the Italian and Latin by Crew, H., De Salvio, A. Introduction by Favaro, A. Original title *Discorsi e dimostrazioni matematiche, intorno a due nuove scienze*, *Discourses and Mathematical Demonstrations Relating to Two New Sciences*
- [Galilei, 1957] Galilei, G.: *The Assayer* (1623). In: Drake, S. (ed. & trans.) *Discoveries and Opinions of Galileo*, pp. 231–280. Doubleday, New York (1957)
- [Gendler, 1998] Gendler, T.S.: Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science* 49(3), 397–424 (1998)
- [Gendler, 2004] Gendler, T.S.: Thought experiments rethought – and reperceived. *Philosophy of Science* 71, 1152–1164 (2004)
- [Giere, 1988] Giere, R.N.: *Explaining Science: a Cognitive Approach*. University of Chicago Press, Chicago (1988)
- [Giere, 2007] Giere, R.: An agent-based conception of models and scientific representation. *Synthese* 172, 269–281 (2007)
- [Giere, 2009] Giere, R.: Why scientific models should not be regarded as works of fiction. In: Suárez, M. (ed.) *Fictions in Science. Philosophical Essays on Modeling and Idealization*, pp. 248–258. Routledge, London (2009)
- [Godfrey-Smith, 2006] Godfrey-Smith, P.: The strategy of model-based science. *Biology and Philosophy* 21, 725–740 (2006)
- [Godfrey-Smith, 2009] Godfrey-Smith, P.: Models and fictions in science. *Philosophical Studies* 143, 101–116 (2009)
- [Hintikka, 1998] Hintikka, J.: What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles S. Peirce Society* 34, 503–533 (1998)
- [Hutchins, 1999] Hutchins, E.: Cognitive artifacts. In: Wilson, R.A., Keil, F.C. (eds.) *Encyclopedia of the Cognitive Sciences*, pp. 126–127. The MIT Press, Cambridge (1999)
- [Kant, 1929] Kant, I.: *Critique of Pure Reason*. MacMillan, London (1929); Translated by Kemp Smith, N. originally published (1787), reprint (1998)
- [Kirsh and Maglio 1994] Kirsh, D., Maglio, P.: On distinguishing epistemic from pragmatic action. *Cognitive Science* 18, 513–549 (1994)
- [Kuorikoski and Lehtinen, 2009] Kuorikoski, J., Lehtinen, A.: Incredible worlds, credible results. *Erkenntnis* 70, 119–131 (2009)
- [Magnani, 2001a] Magnani, L.: *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic/Plenum Publishers, New York (2001)
- [Magnani, 2001b] Magnani, L.: *Philosophy and Geometry. Theoretical and Historical Issues*. Kluwer Academic Publisher, Dordrecht (2001)
- [Magnani, 2004a] Magnani, L.: Conjectures and manipulations. Computational modeling and the extra-theoretical dimension of scientific discovery. *Minds and Machines* 14, 507–537 (2004)
- [Magnani, 2004b] Magnani, L.: Model-based and manipulative abduction in science. *Foundations of science* 9, 219–247 (2004)
- [Magnani, 2007] Magnani, L.: Abduction and chance discovery in science. *International Journal of Knowledge-Based and Intelligent Engineering* 11, 273–279 (2007)
- [Magnani, 2009] Magnani, L.: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Heidelberg (2009)
- [Magnani, 2011] Magnani, L.: *Understanding Violence. The Interwining of Morality, Religion, and Violence: A Philosophical Stance*. Springer, Heidelberg (2011)

- [Mäki, 2009] Mäki, U.: MISSING the world. Models as isolations and credible surrogate systems. *Erkenntnis* 70, 29–43 (2009)
- [Manders, 2008] Manders, K.: The Euclidean diagram. In: Mancosu, P. (ed.) *Philosophy of Mathematical Practice*, pp. 112–183. Clarendon Press, Oxford (2008)
- [Miller, 2010] Miller, G.A.: Mistreating psychology in the decades of brain. *Perspectives on Psychological Science* 5, 716–743 (2010)
- [Mizrahi, 2011] Mizrahi, M.: Idealizations and scientific understanding. *Philosophical Studies* (2011), doi: 10.1007/s11098-011-9716-3
- [Morrison, 2009] Morrison, M.: Fictions, representations, and reality. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 110–135. Routledge, London (2009)
- [Mumma, 2010] Mumma, J.: Proofs, pictures, and Euclid. *Synthese* 175, 255–287 (2010)
- [Nersessian and Chandradekharan, 2009] Nersessian, N.J., Chandradekharan, S.: Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research* 10(3), 178–188 (2009)
- [Nersessian, 1993] Nersessian, N.J.: In the theoretician’s laboratory: thought experimenting as mental modelling. In: Hull, D., Forbes, M., Okruhlik, K. (eds.) *PSA 1992*, East Lansing, MI, vol. 2, pp. 291–301. Philosophy of Science Association (1993)
- [Newton, 1999] Newton, I.: *Philosophiæ Naturalis Principia Mathematica*. General Scholium (1726), 3rd edn. Cohen, I. B., Whitman, A. (trans.) University of California Press, Berkeley (1999)
- [Odling-Smee *et al.*, 2003] Odling-Smee, F.J., Laland, K.N., Feldman, M.W.: *Niche Construction. The Neglected Process in Evolution*. Princeton University Press, Princeton (2003)
- [Park, 2011] Park, W.: Abduction and estimation in animals. *Foundations of Science* (2011), doi: 10.1007/s10699-011-9275-2
- [Peirce, 1931-1958] Peirce, C.S.: *Collected Papers of Charles Sanders Peirce*. In: Hartshorne, C., Weiss, P. (eds.) vol. 1-6; Burks, A.W. (ed.) vol. 7-8. Harvard University Press, Cambridge (1931-1958)
- [Portides, 2007] Portides, D.P.: The relation between idealization and approximation in scientific model construction. *Science & Education* 16, 699–724 (2007)
- [Robinson, 1966] Robinson, A.: *Non-Standard Analysis*. North Holland, Amsterdam (1966)
- [Rouse, 2009] Rouse, J.: Laboratory fictions. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 37–55. Routledge, London (2009)
- [Rowbottom, 2009] Rowbottom, D.P.: Models in biology and physics: what’s the difference. *Foundations of Science* 14, 281–294 (2009)
- [Steel, 2010] Steel, D.: Epistemic values and the argument from inductive risk. *Philosophy of Science* 77, 14–34 (2010)
- [Stjernfelt, 2007] Stjernfelt, F.: *Diagrammatology. An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics*. Springer, Berlin (2007)
- [Suárez, 2009a] Suárez, M.: Scientific fictions as rules of inference. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 158–178. Routledge, London (2009)
- [Suárez, 2009b] Suárez, M. (ed.): *Fictions in Science: Philosophical Essays on Modeling and Idealization*. Routledge, London (2009)
- [Suárez, 2010] Suárez, M.: Fictions, inference, and realism. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 225–245. Philosophia Verlag, Munich (2010)
- [Sugden, 2000] Sugden, R.: Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology* 7, 1–31 (2000)

- [Sugden, 2009] Sugden, R.: Credible worlds, capacities and mechanisms. *Erkenntnis* 70, 3–27 (2009)
- [Thagard, 2002] Thagard, P.: The passionate scientist: emotion in scientific cognition. In: Carruthers, P., Stich, S., Siegal, M. (eds.) *The Cognitive Basis of Science*, pp. 235–250. Cambridge University Press, Cambridge (2002)
- [Thom, 1988] Thom, R.: *Esquisse d'une sémiophysique*. InterEditions, Paris (1988); Meyer, V.(trans.): *Semio Physics: a Sketch*. Addison Wesley, Redwood City (1990)
- [Thomson-Jones, 2010] Thomson-Jones, M.: Missing systems and the face value practice. *Synthese* 172, 283–299 (2010)
- [Toon, 2010] Toon, A.: The ontology of theoretical modelling: models. *Synthese* 172, 301–315 (2010)
- [Vorms, 2010] Vorms, M.: The Theoretician's Gambits: Scientific Representations, Their Formats and Content. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 533–558. Springer, Heidelberg (2010)
- [Weisberg, 2007] Weisberg, M.: Three kinds of idealizations. *Journal of Philosophy* 104(12), 639–659 (2007)
- [Woods and Rosales, 2010a] Woods, J., Rosales, A.: Unifying the fictional. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 345–388. Philosophia Verlag, Munich (2010)
- [Woods and Rosales, 2010b] Woods, J., Rosales, A.: Virtuous Distortion. Abstraction and Idealization in Model-Based Science. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 3–30. Springer, Heidelberg (2010)
- [Woods, 2010] Woods, J. (ed.): *Fictions and Models: New Essays*. Philosophia Verlag, Munich (2010)

# An Examination of the Thesis of Models as Representations

Dachao Li and Ping Li

**Abstract.** This paper aims to discuss four versions of the thesis of models as representations that are used to deal with the problem of scientific representation: models as *structures*, *analogs*, *fictions*, and *mental representations*. In particular, the paper focuses on an examination of the problems for the thesis of models as structure and shows that (i) structure cannot be viewed as the essence of models; (ii) isomorphism cannot define a representational relation; and (iii) models involve linguistic descriptions instead of pure abstract mathematical entities. Based on the conception of models as mental representations, the paper suggests a naturalist approach to scientific representation and a reduction of the problem of scientific representation into the problem of mental representation, by which the representational role of models in science may be explained by means of the representational function of mental representations.

## 1 Introduction

Models play a key role in the organization and acquirement of scientific knowledge, in the processes of theorizing in science, and in the representational and reasoning practice of science in history and of the working scientist. In recent decades, many philosophical debates on the problem of scientific representation presuppose the thesis of models as representations, which views *representing* as a fundamental function of models in science and seems to have become a common pre-theoretic intuition among most contemporary philosophers of science. On the other hand, an appropriate natural way to address the representational role of models and the relationship between models and their target systems is increasingly a focus among those philosophers adopting a naturalistic approach to scientific representation.

This paper will make an attempt to look for such a natural way through an examination of the thesis of models as representations. In order to indicate the

---

Dachao Li · Ping Li

Department of Philosophy, Sun Yat-sen University, Guangzhou, China

e-mail: dachao\_sandy@126.com, hsslip@mail.sysu.edu.cn



significance of such an examination on historical and theoretical background at the very beginning, we will immediately present a brief review of how the thesis became popular as a replacement of the notion of models as instantiations.

## 2 Models as *Instantiations vs. Representations*

In standard philosophy of science, models are of less significance and have, if any, only an auxiliary role in a philosophical analysis of the structure of scientific theories. More specifically, logical positivism argues that a scientific theory contains three parts (Nagel 1961, p. 90): (1) an abstract calculus with some logical architecture that also implicitly defines the basic concepts of the system; (2) a set of corresponding rules that give the calculus some kinds of empirical content and link the calculus and empirical observations; and (3) a model of the abstract calculus, which supposes an interpretation for the calculus with more specific concepts or empirical objects. Among the three parts, the calculus and the corresponding rules are of fundamental importance, but the functions of the model are only to aid theoretical construction and formulation and to prompt new research ways and corresponding rules. These functions are completely based on the fact that models are the convenient appearances of the abstract calculus. In all, according to the logical positivist view, the basic role of models is to exemplify a theory. *Models as instantiations* can suppose a variety of heuristic benefits in consideration of people's cognitive limitations, but they are not necessary in a logical sense so far as their contributions to scientific knowledge itself are concerned. In other words, models do not have a representational role in our scientific knowledge about the world; and they do not function as a kind of content vehicles as do scientific theories and laws.

The huge gap between the received view and scientific practice led to the rise of the semantic view of theories (e.g., Suppes 1960, 2002; van Fraassen 1980; Giere 1988, 2004) which treats a theory as a set of models instead of an abstract calculus. In the semantic view, the traditional relationship between theories and models is turned upside down. This reversal highlights the point that models hold a central position in every field of scientific practices and brought the conception of *models as representations* into fashion in recent decades. According to the semantic view, models are primary expressional media in science even though the form of logical calculus might still be used to describe the fundamental presuppositions of models. The corresponding abstract calculus becomes a way of description in a derived sense; and the relation between an abstract calculus and the empirical world must be illustrated only through a model as representation.

There are many versions of the semantic view sharing the belief that models play a central role in the process of organization and acquisition of scientific knowledge, most of which represent their target systems in some way. The divergence among these versions comes to a large extent from two basic questions about models. The first is the ontological problem: what is model? The second is the representational problem: how does a model represent its target system?

The direct answer is to eliminate those problems (French 2010), changing the original problems to a naturalistic perspective. The important problem is not what

is a model, but the question of how to understand models in order to explain the uses and functions of models in scientific practice. The models used in scientific practice show great heterogeneity. We can understand scientific modeling better through a single representational framework that unifies many discrete elements of modeling. But a unified interpretation of models is not necessarily understood as a judgment on the essence of models.

For the sake of an inquiry into the possibility of a unified interpretation of models, this paper will make an examination of four versions of the thesis of models as representations: models as *structures*, *analogs*, *fictions*, and *mental representations*. On the basis of the thesis of *models as mental representations*, the paper will propose a unified framework of scientific representation from a cognitive view.

### 3 The Problems for the Structural View of Models

There are two main versions of the semantic view of theories that focus on a relation between models and their target systems. One is based on the concept of structure isomorphism, called “*the structural view of models*”; another relies upon the relationship of similarity. We will examine the structural view in this section and then return to the second version and other approaches to models in science in the following sections.

The central claim of the structural view is that a model is an abstract mathematical entity such as a structure. A *structure* may be defined as a composite object consisting of three parts, one of which is a non-empty set of individuals, labeled as the *universe* of the structure; the second is a set of operations defined on the universe; and the third is a set of relations defined on the universe. Many advocates of the semantic view argue for models as abstract mathematical entities: “[T]he meaning of the concept of model is the same in mathematics and the empirical sciences.” (Suppes 1960, p. 289) “A scientific theory gives us a family of models to represent the phenomena... These models are mathematical entities, so all they have is structure, the only thing they can represent is structure.” (van Fraassen 1997, pp. 528-9)

Although there are many discussions on the meaning of *models as structure*, the typical structural view of models can be characterized by the scheme: a model (M) is a structure (S); and M *represents* a target system (T) if and only if there is *isomorphism* between T and S.

It is important to note that, in the above scheme, the conceptions of *models as structures* and *as representations* are the core of the semantic approach to theories that Suppes and van Fraassen advocate. The requirement of *isomorphism* is somewhat abated in some other versions of the semantic view, such as Mundy’s homomorphism (1986). In addition, some semantic views assert that models are ultimately data models, which are the representations of data rather than external objects. Here laying aside the differences between such versions and focusing on the above scheme, we will analyze the problems for the structural view of models and show our conclusions: (i) structure and model cannot be equal; (ii) an isomorphic relationship cannot be used to account for scientific representation; and (iii)

there are questionable presuppositions in asserting the isomorphism between models and their target systems.

The structural view of models implies the equivalence between models and structures. This result leads to a series of conceptual issues and clashes with modeling practices in science.

According to such a relationship of equivalence, the same structure means the same model. Therefore, we cannot differentiate two models if they have one and the same mathematical form. That often conflicts with our intuitions. For example, there is a common mathematical form – a second order constant coefficient differential equation – between the model of mechanical damping vibration and the model of electromagnetic vibration. From the view of the modeler, the structures of the two models are exactly the same; but we will intuitively think that they involve two different models. We can even imagine a situation in which a modeler without any knowledge of electromagnetic vibration in a circuit is studying the mechanical damping vibration model while another modeler without an understanding of mechanical damping vibration is working on a circuit's electromagnetic vibration model. In this case, both of them are treating exactly the same structure but quite different models.

The crux of the identity problem seems to be that the notion of structure as abstract entity is not sufficient to define the concept of model. As a result, we cannot distinguish different models only in light of their structures. A natural way to solve this problem is to include some specific descriptions of a target system as part of its model – This means that models are not pure non-linguistic entities. Certainly, such a revision would destroy a unified account for theories and models that the semantic view of theories supposes.

According to the structural view of models, theoretical models are totally, at least apparently, different from the so-called physical models. This brings us to the second problem for the structural view. In scientific practice, some models are only concrete physical objects representing some phenomena or some aspects of the world, such as scale models, illustrations, and diagrams. Physical models are usually static objects; so it is enough to consider their own properties for specific purposes. In some complex situations, we need to operate certain physical objects. These objects as well as the associated operations form models representing some processes. In other words, these models include important dynamic characteristics, so that those processes have to be represented by means of certain causal properties of the physical objects in the models. Unlike theoretical models or abstract mathematical entities, both static physical models and operation-involved physical models do not result in any intractable ontological problems. They are simply physical objects; and their physical and causal properties may help us to understand the objects that they represent. On the contrary, structures as abstract mathematical entities may not have any causal properties. Thus it seems impossible for the structural view of models to provide a unified conceptual framework of scientific models.

In fact, we can unify physical and non-physical models within a conceptual framework of the semantic approach through an analysis of structure illustrated in physical models. But this presupposes that physical objects have certain structures,

which will be discussed when we consider the problem of the structure of target systems.

The third problem for the structural view of models came from the fact that it tends to favor a philosophical analysis of models as representations in terms of the relationships of isomorphism and reference and of the concept of truth, losing sight of a psychology of scientific discovery and of scientific understanding/explanation. Of course, this is an objection from a naturalist point of view. Why not a naturalized stance or a cognitive view in the contemporary philosophy of science?

In scientific practice, theoretical models often can give us some new intuitions and insights about a field, such as Poincaré's perturbation model in solving a many-body problem, by which we could examine the movement of a small particle influenced by two celestial bodies and obtain an intuitive insight into a new kind of movement. Another example is the black hole model, which had brought us many insights about this possible kind of celestial bodies before we found real black holes. If theoretical models are mathematical objects, we can hardly explain the heuristic functions of models because a purely mathematical object cannot provide any intuitions about new movement forms or unknown celestial bodies regardless of any other functions that it possesses. So there must be other factors that help these models to play a heuristic role. Obviously, the structural view of models misses those factors. Moreover, the scheme of models as structures has been exceeded by an increasing literature on many topics such as model-based reasoning (e.g., Nersessian 2002a, 2002b; Magnani 2004a, 2004b), visualization (e.g., Gooding 2004, 2005, 2010), and mathematical representations (e.g., Tweney 2009, 2011, forthcoming) in science.

Fourthly, it is also difficult to treat the problem of theoretical modeling within the framework of models as structures. We can construct many models, such as classical mechanical models, that do not have a full formulation in a mathematical language. This kind of mathematical descriptions requires (at least) that we have a complete axiomatized classical mechanics, so that we can completely determine the theoretical content embodied in the models and fully clarify their mathematical structures. Since these models can be constructed in the absence of a full linguistic or structural description, the construction process of these models cannot be analyzed through the structural view of models.

The arguments embodied in the above analysis runs roughly as follows: From the thesis of models as structures, in the first place, we can derive a claim that there are no models without a structural description or a full mathematical formulation at all, or that there are no models that can be constructed in the absence of a structural description. In the case of classical mechanical models, this claim means that a complete axiomatized classical mechanics is a requirement for the construction of any models. Accordingly, it is impossible to construct (say) classical mechanical models without a known structure unless we have had a complete axiomatized classical mechanics; otherwise, we can't determine the theoretical content embodied in the models and clarify their mathematical structures. And then it is reasonably argued that we can build many classical mechanical models without a known structure; or it is indeed illustrated that many models can be

constructed in the absence of a structural description. Therefore, we conclude that the structural view of models fails to provide an account for how those models without a known structure can be constructed.

Finally, it is impossible for the structural view of models to account for scientific representation appropriately in terms of a relation of isomorphism between models and their target systems. As a matter of fact, it claims a wrong version of the thesis of models as representations. Firstly, it is obvious that the formal attributes of isomorphism are entirely different from the formal attributes of the representational relation.

According to Goodman's argument (1972) against the similarity view of representation, since a relationship of similarity has the properties of symmetry and self-reflection while a relationship of representation has none of these properties, representational relations cannot be taken as similarity relations regarding their formal properties. Here the symmetry of the similarity relation is that if  $x$  is similar to  $y$ , then  $y$  is similar to  $x$ ; and self-reflection means that each object  $x$  is similar to  $x$  itself. So far as the representational relation is concerned, we can assert at least very plausibly that it does not have the features of symmetry and self-reflection even if we cannot prove that it has exactly the opposite features – that is, if  $x$  represents  $y$ , then  $y$  does not represent  $x$ ; and  $x$  never represents  $x$ . In fact, Goodman's argument can be used to reject the isomorphism-based representational relation on the basis of the fact that the isomorphic relation obviously has the features of symmetry and self-reflection. In short, isomorphism cannot be a representational relation with regard to formal properties.

Secondly, an isomorphic relation is not sufficient to define the representational relation because there are many cases in which there is an isomorphic relation but not a representational relation between two objects, for example, two items manufactured from the same mold. With respect to scientific models, two different models can have the same structure because the structure has the feature of multiple instantiation. In the example mentioned above, the mechanical damping vibration model is isomorphic both to the phenomena of mechanical damping vibration and to electromagnetic vibration models, but it can only represent the former rather than the latter.

These two problems resulted certainly from our attempt to explain a representational relationship purely based on the isomorphic relationship. If we adopt Giere's proposal (2004, 2010) and introduce the scientist's purpose or intention as a necessary component for understanding scientific representation, then the corresponding structural view of models can be modified as follows: a model ( $M$ ) is a structure ( $S$ ), and  $M$  represents a target system ( $T$ ) if and only if  $T$  is isomorphic to  $S$ , and  $M$  is used by  $U$  for some purpose ( $P$ ). The "U" may be "an individual scientist, a scientific group, or a larger scientific community." (Giere 2004, p. 743). In this revision, we can avoid those problems discussed in the previous two paragraphs, but the isomorphism between models and target systems seems to be totally irrelevant to representational relations between both of them. Once resorting to a user's intentions in an interpretation of the representational relationship without any other limitations, every object can represent every other object simply because this is the purpose of the user. Such a general understanding of scientific

representation cannot be appropriate because it cannot account for how we learn from models in science. Nevertheless, we still believe that the user of representations, with his/her purposes and intentions, is an indispensable element for an understanding of scientific representation.

In addition, inaccurate and even entirely wrong models are very commonly found in scientific practice. Many models are based on idealized assumptions. They are simplified representations of target systems. While constructing the models in question, we definitely know most of those assumptions are wrong in the sense that they do not hold in the reality, or that they do not match their target systems accurately. According to the structural view of models, either models are isomorphic to their target systems and thus represent them, or they do not have any representational functions due to the lack of a relationship of isomorphism. Therefore, the structural view completely denies the representational function of inaccurate models. Although some versions of the semantic view of theories relax the isomorphic requirement, we cannot attribute any representational functions to completely wrong models (for example, the perpetual motion model and the ether model) in accordance with the semantic approach. During a scientific revolution there are often important models involving contradictions, such as the Bohr atom model, that contain theoretical elements from both new and old paradigms. For these kinds of models, we cannot explain their structures and attribute to them any representational functions based on the semantic approach. These problems suggest that the structural view of models cannot explain the possibility of misrepresentation. *A theory of scientific representations cannot be plausible if it rules out the possibility of misrepresentation.*

An isomorphic relation holds only between two objects with some structure, so the structural view of models must presuppose that a target system presents the same structure as a model that represents the target system. The target system as a physical object, however, does not show only a unique structure that is to be compared with the structure of the model because the structure presented by the physical object depends on how we choose the individuals constituting the target system and the relations among those individuals. Therefore, the target system can display different and non-isomorphic structures based on the ways we describe it, which in turn depend on the background of particular research work and scientific problems. To assert isomorphism between a model or structure and a target system, we have to presuppose that the target system has a specific structure and then stipulate certain descriptions of the target system. These kinds of descriptions should be a necessary element in any analyses of scientific representation. In other words, we need to give up the core point – that is, models are purely non-linguistic entities – of the semantic approach to the structure of theories in order to explain the representational function of models in science.

## 4 Other Approaches

Giere (1988, 2004) advocated an alternate version of the semantic view of theories, “*the similarity view of models,*” and insisted on the claim that there is a similarity relation rather than isomorphism between models and their target systems.

According to this claim, we can answer the representational problem as follows: a model *M* represents the target system *T* if and only if *M* is similar to *T*. Compared with the structural view of models, Giere's point of view more or less limits the answers to what is a model or what is a scientific representation.

Giere's proposal has some obvious advantages. First, it allows that a model needs only to be similar to its target system, which is next to the commonsense concept of models in science. Second, it does not commit a specific answer to the ontological question and thus allows a variety of physical or abstract objects as models, which is consistent with the great heterogeneity displayed by models in scientific practice.

We think that the similarity view of models does not avoid the above-mentioned problems faced by the structural view with regard to the representational problem, and that the similarity view introduces another serious problem on the other hand. As we know, every object is similar to every other object in some way, so there is no content at all in the assertion that the model and the target system are similar to each other. In order to go ahead, we have to show clearly those aspects in which the model and the target system are similar and the degree of similarity between both of them. The things that can help us to do so may be the so called "theoretical hypotheses" (Giere 1988, p. 81). That is to say, the judgment of similarity must rely on accompanying linguistic descriptions; and this may be viewed as a further reason why we reject the concept of models as mere non-linguistic entities.

Fictionalism (Frigg 2010) is another alternative approach to models in science, which claims that a model is a fictional object, namely, an imaginary concrete object. Here the concept of fictional objects is ambiguous, one sense of which is that the fictional objects are not real at all, and another sense of which is that the fictional objects should be concrete if they would be real. The latter kind of sense seems to be the precise meaning of the concept used by fictionalists.

According to the fictionalist thesis of *models as fictions*, for example, a supply and demand model in economics actually consists of a virtual group of buyers and sellers who are very similar to the actual buyers and sellers. This group of people has many properties attributed in the construction of such a model, such as preferences, commodity, price, and budget. At the same time there are many properties that are not described in the construction of the model, which may be later stated as a supplement or precision in the improvement of the model or derived from the existing features of the model.

This view accords with the way by which scientists talk about models in scientific practice. As a matter of fact, scientists talk about models as if these models are concrete objects, so that more and more new properties that they possess often can be found out through the further study of those models – for example, by way of the derivation of theorems from the assumptions of models.

The fictionalist account of scientific representation has an apparent problem in talking about fictional objects in consideration of Quine's effort (1948) to completely eliminate the references of those imaginary concrete entities. In his opinion, bad grammars would lead us into faulty ontologies. As an illustration, we may make an inference as follows: Pegasus must exist in some sense – because we

have to be talking about something if we say "Pegasus doesn't exist". Quine supposed that we wrongly take asserting "Pegasus doesn't exist" to be ascribing a property (of nonexistence) to an object (Pegasus). This misinterprets that sentence. But we can clearly re-express it as "It is false that there is something that is Pegasus", which doesn't assume that Pegasus in some sense exists. For the sake of avoiding the ontological pitfalls, Frigg (2010) tries to clarify the concept of imaginary objects more clearly with the aid of the camouflage theory that is concerned with imitation in arts.

Granted that the ontological obstacle was excluded, fictionalism still faces some problems. At first, there must be great individual differences among different scientists in imagining model systems if models are conceived as imaginary objects. But the highly consistent properties of a model are a necessary condition for the model to play a role in scientific practice or within a scientific community. Secondly, with respect to a scientific methodology, how do scientists compare the models as fictional objects with the target objects in the real world. It is impossible for scientists to observe/measure fictional entities themselves and/or their attributes since they are mere imaginary objects. Furthermore, the fictionalist theory cannot even supply the identity condition of models. Of course, these challenges to fictionalism apply to only those fictional objects without a full mathematical formulation. Frigg's fictionalist account considers only abstract models as fictions, the most important kind of models in science; but it seems that such fictional entities are hardly regarded as the vehicles of scientific knowledge and as scientific models due to the lack of the identity condition. However, we may view fictionalism as an important supplement of the structural view of models in the sense that the fictionalist theory of scientific representation does offer some useful clues to account for the previously mentioned important elements of models missed by the semantic approach to theories and models in science. In particular, the main roles may be the mental representations of models, instead of the suspicious fictional entities. Thus we will discuss the last version of the representation thesis, *models as mental representations*<sup>1</sup>, in the following section.

## 5 Mental Models and the Elements of Modeling

Models as mental representations may involve many representational forms such as images, schemata, scripts, logical and causal mental models, and so on. Rather

---

<sup>1</sup> One objection that would invalidate this thesis is the question: Would scientific work currently done by robots qualify? So far as the auxiliary work done by robots in scientific practice is concerned, scientists may say that they do undertake a lot of scientific work. However, robots' models can't be strictly taken as ones used by scientists unless we undoubtedly accept the claim that robots currently are thinking by the same way as do scientists. Thus if it is exaggerated to say that robots' models are mental ones, it also seems exaggerated to claim that their thinking is the same as scientists' one. In addition, the term "models as mental representations" here is used in the sense that we accept Giere's intentional conception of representation in science (2010) and put scientists as well as their intentions and purposes in the picture of scientific representation.



than the exact forms or a generic form of models as mental representations within the context of scientific representation, we will take Johnson-Laird's concept of mental model (1983) as a preliminary candidate for an analysis of the mental representation of models in science.

Like logical empiricists, as previously stated, early advocates of the semantic view of theories attempted to construct a unified account of the structure of scientific theories, which actually displayed obvious differences from scientific practice. This situation is consistent with another important trend in the field of philosophy of science: Hansen, Kuhn, and others had rejected the concept of rational justification based on a logical analysis and sought an interpretation of conceptual changes and rational progress in science based on the history of science. Thus the later development of the semantic view paid more attention to the real scientific practices and made an effort to characterize the features of models in scientific practice.

If we firmly follow this tendency, our examination of models in science should proceed by means of the natural context of scientific modeling. Thus the process of modeling is to be analyzed by the basic cognitive processes and mechanisms that constitute the kinds of operations on mental representations of models.

In the semantic view, on the other hand, a theory is not a linguistic entity but an abstract non-linguistic object that can be formalized. This premise implies *a priori* denial of the claim that there is a kind of content that cannot be formulated by a formal language, and then means that any non-formal expressions (e.g., graphics, images, visualizations, analogies, etc.) only can serve a demonstrative or auxiliary role. In fact, logical empiricists had made such judgments on the models in quantum mechanics until it was realized that those models cannot be equivalent to the formalized formulation. This fact is an important support to the semantic view of theories. If we are open-minded for the premise, we should argue for the thesis of models as mental representations and involve mental models in the scheme of modeling in science. The problem of whether there is content that cannot be formulated by formal languages should wait for a solution in future cognitive studies.

The mental representations of models provide a natural way to deal with the representational problem; namely, the representational role of models in scientific thinking is based on the representational function of mental representations. Certainly, the problem of mental representation still is the subject of great controversy and an open question, but we can reduce the problem of scientific representation to the problem of mental representation, so that we can think about the problem from a cognitive view.

Johnson-Laird's conception of mental models can characterize the nature of models as mental representations in science. Mental models have three notable features: (1) structural characteristics; (2) perceptual characteristics; (3) abstract characteristics. Structure characteristics refers to the fact that mental models can contain complex structures; perceptual characteristics shows the fact that mental models share some of the properties of mental images and can be operated upon in the corresponding ways; abstract features mean that a single element of a mental model can represent complicated meanings, such as a complete proposition. These

features also appear in the models in scientific practice and are the base on which models function in science.

The semantic view of theories completely ignores the perceptual features of models in the analyses of non-physical models. Contrary to this trend, Kuhn's theory of scientific revolution attaches importance to this aspect of perception. As an illustration, he pointed out that the puzzle-solving of conventional or normal science and the training processes of entry into a scientific community mainly involve the so-called similarity-grouping process and need to resort to such kinds of perceptual characteristics. And to accept the perceptual features of theoretical models makes it possible to reach a unified understanding of both physical and non-physical models.

Based on the above analyses, we suggest a scheme for understanding theoretical modeling in science that consists of the elements: theory, structure of model, linguistic description of target system, mental model, and target system. And we propose that a model may contain four parts: (i) the description of the model; (ii) the structure of the model; (iii) mental models; and (iv) the description of the target system. The paper will account for the picture of models in brief and contrast it with the structural view of models.

Thus, at first, mental models or the mental representations of models are the representations of potential target systems. The representational role of models is based on the formation of mental representations during the construction of models. On the other hand, it is possible that there are no corresponding target systems at all with respect to some models under study.

Next, theories supply constraints for the descriptions of models. Thus it is possible to clarify the structure of a model through the description of the model. In our scheme of models as mental representations, the relationships among the three elements (namely, theories, descriptions of models, and structure of models) do not conflict with the picture of the structural view of models. Here we simply deny the point that structure is the essence of models.

Moreover, as argued previously, the description of target systems is necessary for analyzing models. This kind of description should involve the individuals constituting a target system and the possible relationships among those individuals, by which we are able to determine the structural type that the target system belongs to. This kind of structure can be isomorphic (or partially isomorphic) to the structure of the model and of the relevant mental models.

Compared with the structural view of models, our proposed framework for understanding models in science has the following characteristics: (1) it is a unified scheme containing both physical and non-physical models; (2) it reduces the problem of scientific representation into the problem of mental representation although it does not resolve the problem; and (3) it changes the status or role of the structure of models in the philosophical understanding of theories and models in science. Structure is no longer the essence of models but one feature presented by models, so that the construction of models does not require a detailed analysis of structure or even a definite structure.

## 6 Conclusion

As a replacement of the notion of models as instantiations, the thesis of models as representations has been widely prevalent for decades and become a pre-theoretic intuition involved in the philosophical debates on the problem of scientific representation. Among various versions of the thesis, this paper has examined four: models as *structures*, *analogs*, *fictions*, and *mental representations*. The analyses of the paper show that the first version, *models as structures*, brings several serious problems that block a natural way to deal with the representational role of models, the relation between models and their target systems, and the picture of model-building appropriately, and that the last version, *models as mental representations*, supports a naturalistic approach to scientific representation, by which the representational role of models in science may be explained on the basis of the representational function of mental representations. The reduction of the problem of scientific representation into the problem of mental representation makes it possible to think about scientific representation from a cognitive view.

However, the paper has presented just an attempt to look for such an appropriate natural way and to inquire into the possibility of a unified interpretation of models in science. Thus it has completed only two main tasks: to include the factors/elements missed by the structural view and suggested by the similarity view and the fictionalist view into our framework of understanding models in science; and to reduce the problem of scientific representation into the problem of mental representation. To have a developed framework of understanding models in science, we have to account for other possible important roles of models and the basic cognitive processes and mechanisms of model-building; and we have to respond other theses of models such as *models as mediators*. All these need some piece of separate work.

**Acknowledgments.** The research work of this paper is supported by the Philosophy and Social Sciences Foundation of the Ministry of Education of P.R. China (Project Number: 11JZD007). The authors really appreciate Professor Ryan D. Tweney for his comments on the draft of the paper and for his helpful suggestions of grammatical corrections. The authors also are grateful to Professor Lorenzo Magnani, Professor Xiang Chen, and the three peer reviewers for their comments.

## References

- Bogen, J., Woodward, J.: Saving the phenomena. *The Philosophical Review* 97, 303–352 (1988)
- French, S.: Keeping quiet on the ontology of models. *Synthese* 172, 231–249 (2010)
- Frigg, R.: Fiction and scientific representation. In: Frigg, R., Hunter, M.C. (eds.) *Beyond Mimesis and Convention: Representation in Art and Science*, pp. 97–138. Springer, Berlin and New York (2010)
- Giere, R.N.: *Explaining Science: A Cognitive Approach*. University of Chicago Press, Chicago (1988)

- Giere, R.N.: How models are used to represent reality. *Philosophy of Science* 71, 742–752 (2004)
- Giere, R.N.: An agent-based conception of models and scientific representation. *Synthese* 172, 269–281 (2010)
- Gooding, D.C.: Cognition, construction and culture: Visual theories in the sciences. *Journal of Cognition and Culture* 4, 551–593 (2004)
- Gooding, D.C.: Seeing the forest for the trees: Visualization, cognition, and scientific inference. In: Gorman, M.E., et al. (eds.) *Scientific and Technological Thinking*, pp. 173–217. Lawrence Erlbaum Associates, Mahwah (2005)
- Gooding, D.C.: Visualizing scientific inference. *Topics in Cognitive Science* 2, 15–35 (2010)
- Goodman, N.: Seven strictures on similarity. In: Goodman, N. (ed.) *Problems and Projects*, pp. 437–446. Hackett Publishing, Indianapolis and New York (1972)
- Johnson-Laird, P.N.: *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge (1983)
- Magnani, L.: Model-based and manipulative abduction in science. *Foundations of Science* 9, 219–247 (2004a)
- Magnani, L.: Reasoning through doing. Epistemic mediators in scientific discovery. *Journal of Applied Logic* 2, 439–450 (2004b)
- Mundy, B.: On the general theory of meaningful representation. *Synthese* 67, 391–437 (1986)
- Nagel, E.: *The Structure of Science. Problems in the Logic of Scientific Explanation*. Harcourt, Brace and World, New York (1961)
- Nersessian, N.J.: The cognitive basis of model-based reasoning in science. In: Carruthers, P., Stich, S., Siegal, M. (eds.) *Cognitive Models of Science*, pp. 133–153. Cambridge University Press, Cambridge (2002a)
- Nersessian, N.J.: Maxwell and “the method of physical analogy”: model-based reasoning, generic abstraction, and conceptual change. In: Malament, D. (ed.) *Essay in the History and Philosophy of Science and Mathematics*, pp. 129–166. Open Court, Chicago and La Salle (2002b)
- Quine, W.V.: On what there is. *The Review of Metaphysics* 2, 21–38 (1948)
- Simon, H.A.: The axiomatization of classical mechanics. *Philosophy of Science* 21, 340–343 (1954)
- Suppes, P.: A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese* 12, 287–301 (1960)
- Suppes, P.: *Representation and Invariance of Scientific Structures*. CSLI Publications, Stanford (2002)
- Tweney, R.D.: Mathematical representations in science: A cognitive–historical case history. *Topics in Cognitive Science* 1, 758–776 (2009)
- Tweney, R.D.: Representing the electromagnetic field: How Maxwell’s mathematics empowered Faraday’s field theory. *Science & Education* 20, 687–700 (2011)
- Tweney, R.D.: On the unreasonable reasonableness of mathematical physics: A cognitive view. In: Proctor, R.W., Capaldi, E.J. (eds.) *Psychology of Science: Implicit and Explicit Processes (to appear)* (forthcoming)
- van Fraassen, B.C.: *The Scientific Image*. Clarendon Press, Oxford (1980)
- van Fraassen, B.C.: Structure and perspective: Philosophical perplexity and paradox. In: Dalla Chiara, M.L., et al. (eds.) *Logic and Scientific Methods*, pp. 511–530. Kluwer Academic Publishers, Dordrecht (1997)

# On Animal Cognition: Before and After the Beast-Machine Controversy

Woosuk Park

**Abstract.** Elsewhere I discussed analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. [Park (2011)] In this paper, I propose to examine the role and function of the Beast-Machine controversy in the fascinating story of the fortuna of animal cognition. For, to say the least, it must be one of the most salient landmarks in the history of the transformation of the problem of animal cognition. In Section 2, I shall briefly recite the analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. In Section 3, after briefly introducing Descartes' infamous hypothesis of animal automatism and the immediate responses to it, I shall focus on three of the most important parties in the controversy, i.e., the Cartesians, the Aristotelians, and the empiricists. My emphasis will be on the argument strategies of each of these parties for their views on intuition and intelligence of animals. In Section 4, I shall discuss why both Cartesians and the empiricist tried to avoid the notion of instinct. Also, I shall argue that the formulation of the issue as between instinct and intelligence itself is a false dilemma, thereby highlighting the greatness of Avicenna's, Peirce's, and Magnani's solutions.

## 1 Introduction

Elsewhere I discussed analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. [Park (2011)]. There are analogies between them (1) at the level of

---

Woosuk Park

Korea Advanced Institute of Science and Technology,

291 Daehak-ro(373-1 Guseong-dong) Yuseong-gu Daejeon 305-701

Republic of Korea

e-mail: [woosukpark@kaist.ac.kr](mailto:woosukpark@kaist.ac.kr)

the problems, (2) at the level of the diagnosis, and (3) at the level of prognosis. Both Avicenna and Peirce-Magnani address their views on the problem of intuition and intelligence of animals. Also, they detect the main cause of the problem in the false dilemma usually posed by the participants of the controversy. Finally, they seek the solution in estimation and abduction in animals respectively that have both intuitive and intelligent elements. On the other hand, some of the disanalogies are also manifest. Unlike Avicenna's estimation, which is one of the internal senses in Aristotelian faculty psychology, abduction in Peirce-Magnani is never such a sense faculty. Further, unlike estimation, which must be ascribed implicitly only to vertebrates, Magnani wants to allow abductive instinct to any kind of organism as well as to vertebrates. Since Avicenna and Peirce-Magnani are a millennium apart, both al these analogies and disanalogies are extremely intriguing. In other words, we need to explain not only how there could be such a remarkable analogies, but also what events in the internal and external history of the problem of animal cognition caused such disanalogies.

In this paper, I propose to examine the role and function of the Beast-Machine controversy in this fascinating story of the fortuna of animal cognition. For, to say the least, it must be one of the most salient landmarks in the history of the transformation of the problem of animal cognition. In Section 2, I shall briefly recite the analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. In Section 3, after briefly introducing Descartes' infamous hypothesis of animal automatism and the immediate responses to it, I shall focus on three of the most important parties in the controversy, i.e., the Cartesians, the Aristotelians, and the empiricists. My emphasis will be on the argument strategies of each of these parties for their views on intuition and intelligence of animals. In Section 4, I shall discuss why both Cartesians and the empiricists tried to avoid the notion of instinct. Also, I shall argue that the formulation of the issue as between instinct and intelligence itself is a false dilemma, thereby highlighting the greatness of Avicenna's and Peirce-Magnani's solutions.

## 2 Between Avicenna and Peirce-Magnani: Estimation and Abduction in Animal

### 2.1 *Avicenna's Sheep and Wolf*

In Avicenna (Ibn Sîna, 980-1037) we find "one of the most complex and sophisticated accounts of the internal senses", which was motivated to expand "Aristotle's notion of imagination or *phantasia*". [Black (1993), 219] In this expansion, the internal sense faculty of estimation (*wahm/aestimatio*) has particular importance "in accounting for features of both animal and human

---

<sup>1</sup> This section is based on Park (2011).

cognition” [Black (2000), 59]<sup>2</sup> The existence of estimation in non-human animals was taken for granted by most medieval philosophers. In what Deborah Black calls “the canonical presentation of estimation” or “the Avicennian paradigm”, typical examples are “the sheeps perception of hostility in the wolf, or its perception of its offspring as an object of love”. [Black (1993), 220]<sup>3</sup> On the other hand, the issue of estimation in humans was mostly neglected in later medieval discussions of internal senses. According to Black,

estimation was viewed primarily as the animal counterpart of the practical intellect, or it was replaced by the cogitative faculty, which in Avicenna’s philosophy had a cognitive function entirely distinct from that of estimation [Black (2000), 59].

From my point of view, what is intriguing in this contrast of estimation in non-human and human animals is that the Latin medieval philosophers seem to have similar difficulties in understanding estimative power as we do in understanding abduction.

Here is a typical passage from Avicenna on a sheep’s perception of the hostility in the wolf:

As for the intention, it is a thing which the soul perceives from the sensed object without its previously having been perceived by the external sense, just as the sheep perceives the intention of the harm in the wolf, which causes it to fear the wolf and to flee from it, without harm having been perceived at all by the external sense.

...

Then there is the estimative faculty located in the far end of the middle ventricle of the brain, which perceives the non-sensible intentions that exist in the individual sensible objects, like the faculty which judges that the wolf is to be avoided and the child is to be loved [Avicenna (1952), pp. 30-31].

Though brief, we can confirm Avicenna’s some of the most famous and influential theses on estimation in non-human animals. Above all, Avicenna makes it clear that the object of estimative faculty is intention. Further, he announces that intention is not something to be perceived by external senses. Black’s excellent exposition of Avicenna’s celebrated example of the sheep’s perception of the intentions of hostility and danger from the wolf is quite helpful for understanding what is going on here:

Hence the sheep cannot literally be said to “smell danger” in the scent of the wolf or “see danger” in the wolf’s eyes, because smell only perceives odors and vision colors and shapes. Rather, concomitant with its seeing and smelling

---

<sup>2</sup> According to Black, a human estimative faculty was posited in addition to the intellect “in order to account for a variety of complex human judgments that are pre-intellectual but more than merely sensible”. Black (2000), 59.

<sup>3</sup> As Black points out, Avicenna uses these examples in almost all his discussions of internal senses. For references to Avicenna’s particular texts, see note 9 of Black (1993), 247.

the wolf, the sheep must perceive these “intentions” of hostility and danger directly through another faculty, its estimative faculty [Black (2000), 60].

Even if we concede that “intention” is an intolerably ambiguous word in medieval philosophy<sup>4</sup>, we do have a very convincing ground to believe that here we have a clue for understanding why Avicenna postulates a seemingly mysterious faculty of estimation as one of the internal senses.

## 2.2 Peirce’s and Magnani’s Poor Chicken

After all these years of extensive discussion, Peircean abduction is still puzzling to us. One of the most pressing issues in understanding abduction is whether it is an instinct or an inference. For many commentators find it paradoxical “that new ideas and hypotheses are products of an instinct (or an insight), and products of an inference at the same time”<sup>5</sup>. As Sami Paavola points out, we seem to face a dilemma: “If abduction relies on instinct, it is not a form of reasoning, and if it is a form of reasoning, it does not rely on instinct” [Paavola (2005), 131].

Fortunately, Lorenzo Magnani’s recent discussion of animal abduction sheds light on both instinctual and inferential character of Peircean abduction<sup>6</sup>. Contrary to many commentators, who find conflicts between abduction as instinct and abduction as inference, he claims that they simply co-exist.

Probably the following text is the most detailed and well-known expression of Peirce’s views:

How was it that man was ever led to entertain that true theory? You cannot say that is happened by chance, because the possible theories, if not strictly innumerable, at any rate exceed a trillion or the third power of a million; and therefore the chances are too overwhelmingly against the single true theory in the twenty or thirty thousand years during which man has been a thinking animal, ever having come into any mans head. Besides, you cannot seriously think that every little chicken, that is hatched, has to rummage through all possible theories until it lights upon the good idea of picking up something and eating it. On the contrary, you think that the chicken has an innate idea of doing this; that is to say, that it can think of this, but has no faculty of thinking anything else. The chicken you say pecks by instinct. But if you are going to think every poor chicken endowed with an innate tendency toward a positive truth, why should you think that to man alone this gift is denied? [Peirce (1931-1958), 5.591; Magnani (2009), pp. 277-278].

<sup>4</sup> For example, in Duns Scotus we an interesting text where various meanings of term “intention” are recognized in the context of explaining the intentionality of light in the medium: (1) an act of the will, (2) the formal reason of a thing, (3) a concept, (4) what ‘intends’ toward the object. [McCarthy: p. 26].

<sup>5</sup> Paavola (2005), 131.

<sup>6</sup> Magnani (2009), especially chapter 5 “Animal Abduction: From Mindless Organisms to Artifactual Mediators”, which was originally published in Magnani and Li (eds.) (2007), pp 3-38.



In his exposition of abductive instinct in non-human and animals, Magnani invokes this paragraph several times as a sure evidence for Peirce's commitment to both non-human and human animal abduction. Peirce's poor chicken seems to be a springboard for Magnani to extend Peircean view of animal abduction even further<sup>7</sup>

In expanding Peirce's views on animal abduction, Magnani seems to assimilate abduction as an instinct and abduction as an inference from both directions.<sup>8</sup> To those who would allow abductive instinct to nonhuman animals but not to humans, he tries to emphasize the instinctual elements in human abductive reasoning. On the other hand, to those who would allow abduction as inference to humans but not to non-human animals, he suggests to broaden the concept of inference, and thereby that of thinking.

For the former project, Magnani cites hypothesis generation in scientific reasoning as a weighty evidence for abductive instinct in humans:

From this Peircean perspective hypothesis generation is a largely instinctual and nonlinguistic endowment of human beings and, of course, also of animals. It is clear that for Peirce abduction is rooted in the instinct and that many basically instinctual-rooted cognitive performances, like emotions, provide examples of abduction available to both human and non-human animals [Magnani (2009), p. 286].

As for the latter project, Magnani wants to secure inferential character of animal abduction from sign activity and semiotic processes found in non-human animals. Here is a lengthy quote from Magnani that makes this point crystal clear:

---

<sup>7</sup> For example, he quotes the following passage from Peirce: "When a chicken first emerges from the shell, it does not try fifty random ways of appeasing its hunger, but within five minutes is picking up food, choosing as it picks, and picking what it aims to pick. That is not reasoning, because it is not done deliberately; but in every respect but that, it is just like abductive inference". [Magnani confers on the article "The proper treatment of hypothesis: a preliminary chapter, toward an examination of Humes argument against miracles, in its logic and in its history" [1901] (in [Peirce, 1966, p. 692]). Another example could be the following discussion: "An example of instinctual (and putatively "unconscious") abduction is given by the case of animal embodied kinesthetic/motor abilities, capable of leading to some appropriate cognitive behavior; Peirce says abduction even takes place when a new born chick picks up the right sort of corn." This is another example, so to say, of spontaneous abduction analogous to the case of some unconscious/embodied abductive processes in humans: [Magnani (2009), p. 276].


<sup>8</sup> This interpretation of Magnani's strategy seems to be supported strongly by his explicit announcement: "I can conclude that instinct vs. inference represents a conflict we can overcome simply by observing that the work of abduction is partly explicable as a biological phenomenon and partly as a more or less 'logical' operation related to 'plastic' cognitive endowments of all organisms" [Magnani(2009), p. 267].

Many forms of thinking, such as imagistic, emphatic, trial and error, and analogical reasoning, and cognitive activities performed through complex bodily skills, appear to be basically model-based and manipulative. They are usually described in terms of living beings that adjust themselves to the environment rather than in terms of beings that acquire information from the environment. In this sense these kinds of thinking would produce responses that do not seem to involve sentential aspects but rather merely “non-inferential” ways of cognition. If we adopt the semiotic perspective above, which does not reduce the term “inference” to its sentential level, but which includes the whole arena of sign activity in the light of Peircean tradition these kinds of thinking promptly appear full, inferential forms of thought. Let me recall that Peirce stated that all thinking is in signs, and signs can be icons, indices, or symbols, and, moreover, all *inference* is a form of sign activity, where the word sign includes “feeling, image, conception, and other representation” [Peirce, 1931-1958, 5.283] [Magnani (2009), p. 288].

## 2.3 The Analogy between Abduction and Estimation

### 2.3.1 Some Analogies

It is interesting to note that some commentators of Aquinas’ writings have interpreted “*vis aestimativa*” as “instinct”. As Judith Barad appropriately points out, some even translated “‘estimative sense’ as ‘instinct’”. [Barad (1995), p. 95] Such an interpretation and translation can be supported to a certain extent by Aquinas’ texts:

For the sheep, seeing the wolf, judges it a thing to be shunned, from a natural and not a free judgment, because it judges, not from reason, but from “natural instinct” [Aquinas, *Summa theologiae*, I 83, 1] 

Interestingly, however, we can also find passages in Aquinas’ writings where he states that “the estimative sense borders on reason” [Aquinas, *On Truth*, 25,2; Barad (1995), p. 95]. So, we seem to have already a minimum ground for setting up an analogy between abduction and estimation at the level of problems:

(The Analogy at the Level of Problem): *Just as there is a controversy over whether abduction is instinct or inference, there was a controversy over whether estimation is merely an instinct or quite akin to reason.*

However, there are in fact more substantial grounds for the analogy. For, as there is a wide spectrum of possible positions toward the problem of abduction in non-human and human animals, there was also an equally wide spectrum of different theories of estimative power in non-human and human

---

<sup>9</sup> English translation in the text is adopted from Barad (1995), p. 95. Original Latin text is as follows: *Judicat enim ovis videns lupum eum esse Fugendum naturali judicio, et non libero, quia non ex collatione sed ex naturali instinctu hoc judicat.*

animals among medieval Islamic and Latin philosophers. The impressive variety of the different diagnosis of the problem and *eo ipso* the different prognosis suggested in turn affected the problem itself. So, there could be an analogy between the evolution of the problem of abduction and that of the problem of estimation.

As we saw above, there is a big difference between Avicenna and Aquinas in their appreciation of instincts. Unlike Avicenna, who allows interesting relations including cooperation between estimation and reason, Aquinas tends to separate instinct and reason rather sharply. If so, someone's answer to the problem of estimation must vary depending on his or her understanding of instinct and reason. Also, we can see that it is extremely difficult, if not impossible, to escape or overcome a well entrenched dichotomies such as sense and intellect. So, Averroes seems to think that insofar as a suggested faculty of estimation is a sense faculty, it must be reducible to external senses. On the other hand, Aquinas seems to think that if there is in humans a sense faculty that has an element of a judgment, it should be at least carefully distinguished from a corresponding sense faculty in non-human animals that works by instinct alone. So, we may suggest the following analogy at the level of diagnosis:

(The Analogy at the Level of Diagnosis): *(1): Just as the controversy about abduction stems largely from different understandings of both instinct and inference, the controversy about estimation originated largely from different preconceptions of both instinct and reason.; (2): Just as there are insurmountable difficulties to count abduction as purely instinctive or purely inferential, there were serious difficulties in treating a judging faculty exclusively at the sensitive level or exclusively intellectual.*

Finally, just as Peirce and Magnani find both instinctual and inferential aspects of abduction, some medieval philosophers found both instinctual and inferential aspects in estimation. At the risk of anachronism, we may understand Avicenna as proposing such a solution in order to avoid the dangers noticed in the analogy at the level of diagnosis. So, we have an analogy at the level of prognosis:

(The Analogy at the Level of Prognosis): *Just as it is promising to resolve the controversy about abduction by allowing both instinctual and inferential character to abduction, some scholastics found a way out of the dilemma for understanding estimation by allowing both characters to estimation.*

### 2.3.2 Some Disanalogies

To the best of my knowledge, there has been no precedent to compare Peircean abduction and estimation in medieval psychology. Given the analogies drawn above, it is rather surprising in some sense that they have escaped anyone's notice. After all, aren't both of them dealing with some remarkable ability of nonhuman animals?

In comparing Magnani's extension of Peircean animal abduction with medieval discussions of estimative faculty, it may not be too difficult to detect several relevant disanalogies between abduction and estimation. One clear difference between Magnani's animal abduction and medieval notions of estimation in animals is this. As we saw above, Magnani seems to ascribe abductive instinct to any organism. On the other hand, the typical examples of the owners of the estimative faculty in medieval psychology are vertebrates. Another, though closely related, difference is found in that while Magnani grants what he calls pseudothought even to extremely lower animals, the owners of the judgmental sense faculty in medieval psychology are again vertebrates. Still another, but again closely related, difference is that unlike Magnani, who interprets any kind of perception as abduction, medieval philosophers count estimation as confined to perceptions of intentions not reducible to external senses and other internal senses.

What is interesting in these disanalogies is that they seem to involve disciplinary debates. On the one hand, we are bound to consider the boundary between psychology and biology. In medieval psychology, estimative faculty was postulated to explain the surprisingly distinguished ability of non-human animals. Here, this remarkable faculty must be ascribed implicitly only to vertebrates. On the other hand, Magnani wants to allow abductive instinct not only to vertebrates but also to any kind of organism. This contrast is striking enough to suggest one fundamental issue: where to locate abductive instinct, in some higher cognitive faculty or in any lower biological function? In other words, there is an important disanalogy at the level of the scope of the abductive and estimative power. On the other hand, it is also quite mandatory to review the history of the relationship between logic and psychology. In other words, there is another apparent disanalogy between Peircean abduction and estimation that might preclude any attempt to draw any analogy between them to get off the ground. Unlike estimation, which is one of the internal senses in Aristotelian faculty psychology, Peircean abduction is never such a sense faculty. After all, abduction is meant to be on a par with deduction and induction, as a genuine logical inference.

### 3 The Beast-Machine Controversy

Our discussion above clearly indicates that a simple perusal of history of psychology should shed light on somewhat neglected aspects of Peirce's thought as well as the old problem of understanding both abductive instinct and abductive inference. But it turns out to be a mere wishful thinking to gather pertinent information from the standard history of psychology. For, it is a mystery that not just the estimation but also the entire category of internal senses seems disappeared in modern philosophical psychology. In view of the fact that contemporary psychologists find the birth date of their discipline only as late as 1879, it is an arduous task to reconstruct the history of internal senses. Where have all the internal senses gone?

I think that every circumstance is pointing to Descartes' infamous doctrine of animal automatism and the subsequent debates related to beast-machine and man-machine controversies in the 18th and the 19th centuries<sup>10</sup>. As portrayed vividly by Rosenfield, the battle on the beast-machine "was waged furiously among philosophers and poets, scientists and theologians, in halls of learning and in the salons of the aristocratic world". [Rosenfield (1940, 1968), p. xxvii] Among the critics of Cartesians, she lumps together dualists, Aristotelians, Neo-Platonists, and Eclectics as traditionalists, and Epicureans and freethinkers as empiricists. [Cf. Rosenfield (1940, 1968), pp. 73-153] Virtually the same map appears from Richards, who focuses on Cartesians, Aristotelians, and the Sensationalists as the major players in the controversy. [Richards (1979), (1981)] Let us briefly scheme each of the three positions focusing on the problem of animal instinct.

### 3.1 *Cartesian Denial of Animal Soul*

According to Rosenfield, we may find the germ of a concept of animal automatism from Descartes' early notebook around 1620. [Rosenfield (1940, 1968), p. 3] Also, as she points out, it is clear that in 1637 when *Discourse on the Method* was published Descartes had definite interest in animal automatism.

It is also a very remarkable fact that although many animals show more skill than we do in some of their actions, yet the same animals show none at all in many others; so what they do better does not prove that they have any intelligence, for if it did then they would have more intelligence than any of us and would excel us in everything. It proves rather that they have no intelligence at all, and that it is nature which acts in them according to the disposition of their organs. In the same way a clock, consisting only of wheels and springs, can count the hours and measure time more accurately than we can with all our wisdom. [CSM, vol. 1, p. 141]

Recently John Cottingham tried to defend Descartes against the standard interpretation by pointing out the vagueness and ambiguity of the so-called beast-machine doctrine of Descartes. In order to deny that Descartes held the monstrous thesis that animals are totally without feeling, Cottingham requested us to consider the following propositions:

- (1) Animals are machines
- (2) Animals are automata
- (3) Animals do not think

---

<sup>10</sup> There is huge literature on Descartes' automaton theory and its aftermath. A nice starting point could be Rosenfield (1940, 1968). Clarke still counts it as the standard account [Clarke (2003), p. 77]. See also Radner and Radner (1996), Sepper (1989) and Sterrett (2002). According to Rosenfeld, "by 1737 if not earlier, the Cartesian defenders of automatism were so thoroughly beaten as to acknowledge defeat" [Rosenfeld (1940, 1968), p. 65].

- (4) Animals have no language
- (5) Animals have no self-consciousness
- (6) Animals have no consciousness
- (7) Animals are totally without feeling.

According to Cottingham, Descartes indeed held theses (1) to (5). However, he believes that there is no evidence that Descartes held (7). [Cottingham (1978), 551-2]

However, Cottingham's attempt to save Descartes has been criticized by many commentators. For example, Peter Harrison criticized Cottingham for his omission of "(1') Animals have no rational soul" from his list. For, according to Harrington, "although (2) does not entail (7), both (2) and (7) are entailed by (1')"<sup>11</sup> Harrington expressed a surprise for this omission, because he believes that "it is the key to understanding how the various claims which Descartes makes about animal minds are linked together" [Harrington (1992), 223]. I would concur with Harrington's assessment of Harrington's attempt. But I would also note that (1') is not ad hoc for Harrington's purpose but at least equally ascribed to Descartes as (7) ever since the beast-machine controversy started. In short, we may safely assume that Descartes ignited the beast-machine controversy with his radical denial of soul and feeling to animals.

According to Rosenfield, "[o]nly after Descartes' death did his thesis of automatism gather momentum, slowly at first, then with increasing velocity". [Rosenfield (1940, 1968), p. 64] However, serious debates incorporating some of the most typical and weighty points already started in the *Objections and Replies* attached to *Meditations*. For example, in the fourth set of *Objections*, Arnauld raised the following criticism:

As far as the souls of the brutes are concerned, M. Descartes elsewhere suggests clearly enough that they have none. All they have is a body which is constructed in a particular manner, made up of various organs in such a way that all the operations which we observe can be produced in it and by means of it\* [\*Discourse, part 5: AT 6:55ff; CSM 1:139ff].

But I fear that this view will not succeed in finding acceptance in people's minds unless it is supported by very solid arguments. For at first sight it seems incredible that it can come about, without the assistance of any soul, that the light reflected from the body of a wolf onto the eyes of a sheep should move the minute fibres of the optic nerves, and that on reaching the brain this motion should spread the animal spirits throughout the nerves in the manner necessary to precipitate the sheep's flight [AT 7:205; CSM 2:144].

Descartes' reply to Arnauld's objection is interesting. For he seems to exploit, probably unduly, the fact that there are even in humans involuntary movements:

---

<sup>11</sup> Harrington points out that "even if we concede that (2), (3) and (6) do not necessarily entail (7), it is clear that for Descartes at least (1') logically necessitates all of (2), (3), (6) and (7)" [Harrington (1979), 225].

Now a very large number of the motions occurring inside us do not depend in any way on the mind. These include heartbeat, digestion, nutrition, respiration when we are asleep, and also such waking actions as walking, singing and the like, when these occur without the mind attending to them. Why should we be so amazed that the “light reflected from the body of a wolf onto the eyes of a sheep” should equally be capable of arousing the movements of flight in the sheep? [AT 7:230; CSM 2:161].

Rosenfield claims that Cartesians in general add to Descartes’ doctrine of animal automatism “a more detailed explanation of the role of the animal spirits in the training of animals”. [Rosenfield (1940, 1968), p. 36] For example, Henricus Regius (Henri de Roy) refutes the empiricist idea that animal training is indicative of animal intelligence by pointing out that “[f]ountains and clocks are after all similarly ‘trained’!” [Rosenfield (1940, 1968), p. 30] Rosenfield also explains how Graud de Cordemoy elaborated a mechanical explanation for wonderful instinct of animals:

What causes sounds to issue from beasts, which have no soul, is simply this: the heart-beat causes the circulation of the blood, which in turn induces the flow of animal spirits. The animal brain is composed of a substance peculiarly sensitive to impressions. When, for instance, a wolf howls, the noise sets in motion the animal spirits assembled, let us say, at the ear of a near-by sheep. These jostle the brain with their vibrations and set into circulation a new flow of animal spirits to the muscles, putting the sheep to flight. The same vibrations which make the beast run cause it to emit cries. In fact, as soon as a sound reaches the creature’s ears, his muscles of articulation are already disposing themselves to produce a similar sound. Animals thus learn their calls and cries from their fellows. This is all a purely involuntary or mechanical process [Rosenfield (1940, 1968), pp. 38-9].

However, it is hard not to confess that all these Cartesian non-theological arguments are rather weak and unconvincing. That is especially so, in view of the fact that Descartes himself claimed that he “proved it by very strong arguments” [AT 7:426; CSM 2:287] According to the Radners, Descartes was referring to the so-called two tests of thought: (1) language test, and (2) the action test. But they criticize Descartes as follows:

In order to keep animals from passing the tests, he has to interpret “thought” in the narrow sense of pure thought or reason. To conclude that animals lack thought in the wider sense, he needs the additional premise that all modes of thinking, including sensations and feelings, presuppose pure thought as a necessary condition. This premise we found to be unwarranted by Cartesian principles [Radner and Radner (1996), p. 79].

They further note that Descartes was more cautious on animals in his correspondence with Henry More. [Cohen (1936)] There Descartes acknowledged the need for a posteriori investigation of animal behavior. [AT 7:358; CSM 2:248] Also, they enumerate some other arguments of Descartes mentioned in his letter to More. One of them, which refers to immortal souls, is clearly

theological. And the Radners claim that even though some such theological concern is not prominent in Descartes himself, “Descartes’s followers took the hint and came up with a set of theological arguments to support the doctrine of animal automatism”. [Radner and Radner (1996), p. 80]

### 3.2 *Aristotelians’ Attack against the Animal Automatism*

It is quite understandable that it was the Aristotelians who most vehemently responded to the doctrine of animal automatism. For this doctrine was meant to destroy their system at its roots. If animal automatism prevails, there would be no room left for the Peripatetic substantial material form, “a substance intermediate between matter and spirit, capable of sensation but not reflection”. [Rosenfield (1940, 1968), p. 80] According to Rosenfield, these most formidable opponents of the Cartesian doctrine were still surviving even in the first half of the eighteenth century. Rosenfield counts Pardies, Daniel, and Regnault, who were Jesuit fathers, as the leading figures in the Aristotelian campaign against the doctrine of animal automatism.

As a sample of Aristotelian hostile attitude, we may invoke Gabriel Daniel, who “was the self-appointed adversary of Ren Descartes”. [Rosenfield (1940, 1968), p. 86]<sup>12</sup> According to Rosenfield, Daniel counted the Cartesian doctrine of animal automatism “as the very touchstone of Cartesianism”. [Ibid.] Even though Rosenfield thinks that such a claim was “obviously an exaggeration”, she is also aptly pointing out that “Cartesianism and animal automatism were integrally associated in the public mind”. [Rosenfield (1940, 1968), p. 87]

Be that as it may, Daniel seems to take a very clever strategy. For as Rosenfield reports,

Father Daniel pricks the automatists’ bubble by conceding that some animal movements are mechanical and denying that all must necessarily be so [Rosenfield (1940, 1968), p. 87].

Let us examine some of the criticisms of Father Daniel against Cartesians based on Rosenfield’s report. First, Daniel sharply detects a fallacy in Descartes’ argument appealing to the involuntary processes in man. For those involuntary processes cannot be “evidence of mechanical nature of every animal” but merely “the likelihood of *some* mechanical movements among animals” [Rosenfield (1940, 1968), pp. 87-88]. Whether Daniel is right in this criticism or not, it is at least quite helpful in understanding and assessing Descartes’ argument. Probably encouraged by the effectiveness of analyzing

<sup>12</sup> Rosenfield also notes that “Pierre Bayle went so far as to term Gabriel Daniel ‘the author who has best refuted Mr. Des Cartes on the soul of beasts’” [Rosenfield (1940, 1968), p. 90]. Here, she is referring to Bayle (1991), “Rosarius, G.”, p. 231.



Descartes' arguments in terms of the standards of evaluating analogical inferences, Daniel seems to generalize his point into a serious criticism, i.e., that Cartesian reasoning is inconsistent:

If it were, he would include men other than himself among the automata. Ah, but other men talk as I do, answers the Cartesian. Yes, but beasts converse, responds his opponent, and you Cartesians still refuse them reason. The fact that one does not understand a language is no excuse for denying its existence or rationality. Since it is only by analogy that we infer the rationality of men other than ourselves, a similar principle should persuade us that beasts are not puppets [Rosenfield (1940, 1968), p. 88].

Daniel seems a formidable disputant quick to identify lacuna in the Cartesian arguments in the debate. For, example, he pointed out that Cartesians "simply guess that since God is all-powerful he could have created the secret springs of the beast-machine". From my point of view, however, the most weighty argument in Daniel is found in his (in a certain sense) empirical evidence:

Take the example of the horse driven to the edge of a precipice, from the bottom of which arises the odor of hay and oats. The steed's obstinate retreat from the edge of the pit cannot be due to the functioning of the bodily machine, which should have propelled him toward the food.

In broader context, Father Daniel's importance lies in the fact that his *Voyage du monde de Descartes* played a significant role in the vivisection controversy. As the practice of vivisection became increasingly widespread in medical circles, the anti-vivisection movement was also quite popular. As Rosenfield aptly pins down, the humanitarian aspect of the Aristotelian concept of animal soul had certain additional sentimental appeal: "If beasts have soul, then it is of course wicked to vivisect, hunt or even eat them" [Rosenfield (1940, 1968), p. 89].

### 3.3 *Empiricists' Double Strategy*

In both Rosenfield's and Richards' interpretations, the ultimate winner of the beast-machine controversy seems to be the empiricists or the sensationalist. However, there seem to be subtle differences in their emphases. According to Rosenfield's interpretation, Aristotelians and the empiricists are the allies in their common battle against the Cartesian beast-machine. On the other hand, in Richards' interpretation, the sensationalists' attack on both Cartesians and Aristotelians is strongly emphasized. That means, there is room for understanding exactly how such a final victory was won by the empiricists.

Rosenfield divides those empiricists involved in the beast-machine controversy into two groups: Epicureans and the freethinkers. Other than Gassendi, who is the foremost figure in the empiricist tradition, Rosenfield discusses Savinien Cyrano de Bergerac, Martin Cureau de La Chambre rather in detail

among the Epicureans. Pierre Bayle is presented as “[t]he immediate ancestor of the eighteenth-century freethinkers”. [Rosenfield (1940, 1968), p. 121] Among the freethinkers, Rosenfield ascribes “a signal role” to Francois Marie Arouet de Voltaire. [p. 128] But the name of Bernard le Bovier de Fontenelle is (together with the names of Bayle and Voltaire) “still surrounded by an aura of fame”. [p. 132]. On the other hand, in Richards’ list of sensationists, who influenced Erasmus Darwin, Charles Darwin and other evolutionists, we find Gassendi, Martin Cureau de La Chambre, Jean-Antoine Guer, The Abb de Condillac, Georges Leclerc, Comte de Buffon, Julien Offray de La Mettrie, Charles-Georges Le Roy. [Richards (1987), pp. 20-31] Of course, there is huge overlap between Rosenfield’s and Richards’ lists, it is somewhat shocking that Richards omits Bayle and Voltaire in the list.

Be that as it may, in both Rosenfield and Richards, “Gassendi’s thesis that animal and man differ only in degrees” is one of the most important factors in the formation of empirical attitude toward animals. In fact, he is at least equally important as Descartes in leading the modern schools of psychology [Rosenfield (1940, 1968), pp. 110-1]. In his fifth Objections to Descartes’ Meditations, he holds that although beasts do not reason as well as men, yet they reason, and there is no difference between their intelligence and ours except in degrees [Rosenfield (1940,1968), p. 112]. According to Gassendi, the human soul is twofold — spiritual and corporeal or sensitive. The latter he identifies with animal soul. “In line with the Democritean atomism, he pictured the animal soul as a composite of material atoms, so fine as to be indiscernible to the eye” [Rosenfield (1940,1968), p. 112].

### 3.3.1 Rosenfield’s Interpretation

In her relatively well balanced report of the controversy, Rosenfield counts Pierre Bayle, who was a celebrated French journalist with Protestant background, as one of the most influential figures in the beast-machine controversy. According to Rosenfield, “[l]ike Voltaire after him, he applied the sharp edge of his critical method to the Peripatetic, Cartesian, and Leibnizian hypotheses, so also to the thesis that beasts have spiritual soul”. [Rosenfield (1940, 1968), p. 122]<sup>13</sup>

Bayle claims that Father Daniel raised very great difficulties against the Cartesians. Further, among those difficulties, “those concerning the soul of beasts considered as machines are the best that could be proposed” [Bayle (1991), p. 231]. According to Bayle’s assessment, Daniel cleverly exploits the unfortunate consequences that can be drawn from the Cartesian paradox:

“[F]or he shows that the arguments of the Cartesians lead us to judge that other men are machines”.

<sup>13</sup> Rosenfield praises Bayle’s writings as “an ever-so-carefully documented history up to his day” [Rosenfield 1940. 1968], p. 122].

Bayle elaborates this point rather extensively, for he believes this consequence of the Cartesian doctrine to be most upsetting:

The Cartesian has no sooner overthrown, ruined, annihilated the view of the Scholastics concerning the soul of beasts, than he finds out that he can be attacked with his own weapons and can be shown that he has proven too much; and that if he reasons logically, he will give up views that he cannot give up without becoming an object of ridicule and without admitting the most glaring absurdities. For where is the man who would dare to say that he is the only one who thinks and that everybody else is a machine? [p. 232].

Rosenfield finds Bayle as criticizing the Cartesian doctrine because “it is contrary to the facts”:

In his observation of animal conduct, Bayle perceives immediately that animals show evidence of intelligence. If one denies it to them, why should one assume that men other than oneself reason? [Rosenfield (1940, 1968), p. 124].

I rather doubt Bayle’s *reductio ad absurdum* argument against the Cartesians can be a proof that “Bayle perceives immediately that animals show evidence of intelligence”. However, such a reading is consonant with Rosenfield’s interpretation, according to which the Aristotelians and the empiricists are the allies sharing certain empirical evidences against the Cartesians.

Even though Bayle admits that Daniel “puts most embarrassing questions to the Cartesians”, he not only criticizes Aristotelians as lacking persuasiveness but also mocks “the Scholastic notion of animal soul as an intermediate substance” [pp. 234-5; Rosenfield (1940, 1968), p. 125]. In order to show the inadequacy of Father Daniel’s position, Bayle claims that it does not satisfy the desiderata for a system we need. What we need, according to Bayle, is a system (1) “that establishes the mortality of the soul of beast”, (2) that establishes a difference in species between the soul of man and that of beasts”, and (3) “that explains the astonishing activities of bees, dogs, monkeys, elephants”. But he argues that none of these three conditions can be satisfied<sup>14</sup>

According to Rosenfield, based on Cartesian dichotomy, i.e., that any object must be either extended or non-extended, Bayle criticized Aristotelian’s appeal to animal sagacity as follows:

The Peripatetics make reflection the criterion of human rationality. But ‘every being that has sense knows that it has it,’ answers Bayle, ‘and all the acts of the sensitive faculty are ... reflexive upon themselves’. In brief, the neo-Scholastics actually destroy any distinction between human and animal soul [Rosenfield (1940, 1968), p. 123].

### 3.3.2 Richards’ interpretation

Unlike Rosenfield, Richards seems a bit biased in his description of the controversy in favor of what he calls the sensationalist position. It must be due

<sup>14</sup> Bayle’s arguments that these conditions cannot be satisfied are too subtle to be assessed here.

to the fact that he was not primarily interested in the beast-machine controversy per se, but understanding it as a background to the problem of instinct and intelligence of animals in natural theology and the evolutionary theories.

Be that as it may, unlike Rosenfield, Richards hardly mentions the fact the Aristotelians and the sensationalists were allies insofar as they fought a common battle with the Cartesian animal automatism. Rather, he is preoccupied with the idea of lumping Aristotelians and Cartesians together as the target for sensationalists. For example, he writes:

Among the Aristotelians, for instance, were the Jesuit Fathers at Coimbra, Portugal, who wrote on instinct during the late sixteenth century<sup>15</sup> They explained instinctive behavior, such as the lambs flight from the wolf, as a function of the animal soul, in which the Creator had instilled sets of behavioral determinants for the preservation of the individual and its progeny. Descartes (1596-1650) and his later disciples, by contrast, denied that animals had substantial souls and that their actions were the result even of primitive cognition. Animals were only machines; their actions came, as it were, wired-in.<sup>16</sup> Both Aristotelians and Cartesians refused to see in animal behavior the least stirrings of intelligence or reason. The action of beasts, they believed, were compelled by blind instinct [Richards (1979), 86].

Richards's serious concern to lump together Aristotelians and Cartesians as the target of sensationalists can be witnessed by more examples. Immediately after having mentioned Reimarus' contention that "the brute was not a mere machine, but from birth harbored in its soul 'an idea or image (*Denkbild*) as a guide and a plan for works and activities of this kind'", Descartes' reference to "'images' and 'ideas' of animal corporeal imagination", and Thomas Willis' description of "the cerebral dispositions determining animal instinct" as 'inborn notion' (*notitia ingenita*), Richards writes as follows:

The theory that animals were possessed of congenital images or notions, even if different from man's, aroused the sensationalist's own instinctive aversion to innate ideas. If all ideas came from sensation, then instincts, as commonly interpreted, could have no existence [Richards (1979), 95].

Of course, Richards is alert to note that "Aristotelians and Cartesians also had hard empirical evidence on their side". [Ibid.] It is quite interesting (and

<sup>15</sup> "An brutae animantes solo natura instinctus in fines suos ferantur" and "Quidnam sit brutorum animantius instinctus," *Commentariorum Collegii Conimbricensis Societatis Iesu, In octo libros Physicorum Aristotelis Stagirita, prima pars, II, ix, quest. 3 et quest. 4* (Cologne, 1602), col. 420-29. (Richards' footnote)

<sup>16</sup> See, for example, Descartes' letter to the Marquess of Newcastle (1646), *Oeuvres de Descartes*, edited by C. Adam and P. Tannery (13 vols; Paris, 1897-1913), IV, 573-75. Thomas Willis's *De anima brutorum quae hominis vitalis ac sensitive est* (1672), in *Thomas Willis Opera omnia*, ed. G. Blasius (Amsterdam, 1862), and Antoine Dilly's *Traitt de l'ame et de la connoissance des btes* (1676) rev. ed. (Amsterdam, 1691) developed and refined Descartes' theory of the beast machine. (Richards' footnote)

potentially important) to point out that Richards here again lumps together Aristotelians and Cartesians, even though he cites only Reimarus (1694-1768) as an example:

Reimarus, taking aim particularly at Gassendi, La Mettrie, Condillac, and Guer, and their attribution of reason to animals, pressed the fact in his *Allgemeine Betrachtunhen ber die Triebe der Thiere* (1760) that animals often exhibited completely formed and adaptive behavior prior to any opportunity for learning from experience: chicks on first emerging from the egg began to peck at seeds with coordinated movements; caterpillars which had never before seen a cocoon, skillfully wove the same patterns as their ancestors. These “skill-drives are executed from birth on, without any experience, instruction, or example, and without error; and thus certainly they are naturally innate and hereditary”<sup>17</sup> [Richards (1979), 95].

Since Reimarus is cited as a neo-Aristotelian on the same page, Richards probably wants to highlight the Cartesian element in Reimarus based on the innatedness of skill-drive. To say the least, however, it could be controversial to lump Aristotelians and Cartesians together based on just one rare example. In fact, Richards, in the parallel text in a later writing, changes “Aristotelians and Cartesians” to “[m]ore traditional theorists of instinct” in the sentence at issue. [Richards (1987), p. 30].

### 3.3.3 A New Interpretation

As we saw above, the sensationalists scorned the use of the term “instinct”. For example, “Guer disdained the concept of instinct as irretrievably joined to an outmoded philosophy”. [Richards (1979), 95] According to Rosenfield, Fontenelle also “ridiculed the notion of instinct” [Rosenfield (1940, 1968), p. 126]. So, it is interesting to note that it was not the sensationalists alone that scorned the use of the term “instinct”. For “Cartesians scorned the use of the word instinct” [Diamond (1974)]. But why should the use of the term “instinct” be scorned by both the sensationalists and the Cartesians? According to Diamond, it was scorned by the Cartesians “because it was then still a motivational term, and machines cannot be said to be motivated”. [Diamond, *ibid.*] In other words, Cartesians’s scorn toward the use of the term “instinct” merely betrays the fact that it was extremely difficult for them to explain instinct in mechanical ways. The sensationalists’ scorn to the use of the term “instinct” was also due to a certain difficulty for them. According to Richards, instincts were the serious problem for the empiricists, because “[i]f all ideas came from sensation, then instincts, as commonly interpreted, could have no existence” [Richards (1979), 95]. In short, both Cartesians and the empiricists appear to scorn or disdain the concept of instinct in order to avoid it, if possible. In the same vein, it is interesting to note that both

<sup>17</sup> H. Reimarus, *Allgemeine Betrachtunhen ber die Triebe der Thiere* (1760), XCIII (3rd ed.; Hamburg, 1773), p. 160. (Richards’ footnote)

the Cartesians and the empiricists are so fond of characterizing instinct as blind or wired. By making instincts as blind or wired, what is beneficial to the Cartesians and the empiricists? How was the project of making instincts blind or wired related to the project of avoiding the problem of instinct?

In order to understand the situation, we need to consider another agenda that has not been brought up so far. It is related to the issue of the continuity of non-human animals and humans in cognition. One hint is available already. As Rosenfield aptly notes, Cartesians exploited extensively the idea that “if we granted animals reason of their own we should have to concede that their intelligence is greater than ours” [Rosenfield (1940, 1968), p. 44]. Now, interestingly enough, the empiricists seem to adopt such an argument strategy in their fight against the Aristotelians. If we are on the right track here, then both Cartesians and the empiricists took advantage of the Achilles heel of the Aristotelians. Aristotelians should preserve the difference in kind between non-human animal cognition and human cognition. Roughly speaking, non-human animals can have senses but no intellectual knowledge. In other words, non-human animals cannot have any knowledge about universals or abstract entities but particulars. Now, if we grant reason of their own to non-human animals by allowing estimation as a judging faculty to them, we might concede that non-human animals are superior to humans in terms of intelligence. Probably such an argument would not have impressed Avicenna. But it would work to Aquinas, who also felt the need to introduce cogitative faculty in humans as a counterpart to estimation in animals. Suppose that the Aristotelians were succumbing to this tricky and rhetorical argument of Cartesians. Since there is no longer estimation, wonderful instincts of non-human animals, it becomes much easier for the empiricists to push their point that animal cognition and human cognition are continuous. There is no cognition whatsoever that is not from the senses.

Now we can notice that both Cartesians and empiricists wanted to avoid not just the term “instinct” but also the “animal soul”. For example, Rosenfield writes:

Voltaire, like Descartes before him, prefers to avoid if possible the term ‘animal soul’. As a general rule he means by it the ‘organization of the body’. Following Bayle’s footsteps, he points out that animal sensitivity, like animal will, is a function of the organization of the body, which idea also explains the difference of degree between man and beast. In a word, animals are machines which think to a greater or lesser degree according to their organization [Rosenfield (1940, 1968), pp. 130-1].

Of course, the empiricists did not call the sensitive soul in animals an intermediary substance or substantial material form. Indeed their mockery of this Peripatetic notion far exceeded in unsparing pointedness anything from the pen of the Cartesians. Owing to the Cartesian attack, Peripateticism was not so formidable by the time the empiricists grew bold. Nevertheless, the traditionalists and empiricists were alike in conceding some measure of mental powers to beasts [Rosenfield (1940, 1968), p. 109].

But here the empiricists' attitude toward animal soul seems somewhat strange. Insofar as empiricists side with Aristotelians in their campaign against the Cartesian hypothesis of animal automatism, thereby allowing souls to nonhuman animals in some sense, it is debatable whether the empiricists' theory of animal soul is superior to that of the Aristotelians. However, it seems as if the empiricist never felt such a need to demonstrate their superiority over Aristotelians. Why? Probably, because Aristotelians were almost destroyed by the Cartesian attack, at least in physics, physiology, and psychology. According to this perspective, the empiricists turn out to be free riders indebted to the Cartesians in destroying the Aristotelian internal senses.

#### 4 Instinct or Intelligence: A False Dilemma

What becomes clear from the discussion above is an intriguing fact that Avicenna's notion of estimation disappeared completely during the period of the beast-machine controversy. It is by no means a small matter. For the medieval Arabic and Latin discussions of animal instinct and intellect were, in some sense, motivated, framed, and developed around that notion. During the 17th century and early 18th century, as we saw above, there were continued discussions of animal instinct and intelligence. Again, as we saw above, Avicenna's sheep and wolf were still there at the center of those discussions. But Avicenna's notion of estimative power of animals was no longer seriously considered. How are we to understand all this?

As was pointed out in the previous section, talk of animal soul itself lost momentum. There is no doubt that the Cartesian doctrine of animal automatism is to blame for the disappearance of animal souls. However, we should not forget the fact, insofar as the Cartesian beast-machine was the target, Aristotelians and the empiricists were allies. For what reasons did the empiricists have to detest or criticize the Aristotelian animal souls? Also, even if they had sufficient reasons for doing that, whatever they may be, did they do the right thing on justifiable ground?

There are also closely related mysteries. Largely speaking, the Aristotelian psychology of internal senses was gone with animal souls. Nevertheless, there seems to be the remnant of internal senses in the British empiricists' writings. For example, we find internal senses of beauty and virtues in Francis Hutcheson. [Hutcheson (1783)] Or, we may cite Lockean introspection as an internal sense in disguise. However feeble, there seems to be the empiricists' appropriation of Aristotelian internal senses. In some sense, all the internal senses in Aristotelian psychology are still with us: common sense, memory, imagination, and instinct. Again, how are we to understand all this?

In order to answer this question, from our discussion in the previous section, one might surmise that there was a conspiracy between the Cartesians and the empiricists to kill the Aristotelian animal souls. Anyway, they had

the common agenda of overcoming the old philosophy. Even though it would be an arduous task to prove that there was indeed such a conspiracy, it is at least possible to evidence that the empiricists were strongly influenced by the Cartesian animal automatism, or even that the empiricists were to a certain extent Cartesians themselves.

Let us take a look at the following quotes from Richards:

The epistemology of sensationalism seemed to be confirmed by the successes of experimental methods in the various sciences and technologies; but toward the end of the eighteenth century, careful observations of animal behavior began to undermine the assumptions of sensationalist epistemology. Naturalists committed to sensationalism thus faced a critical problem, which had implications for their conception not only of animal psychology, but of human psychology as well. The disputes over animal abilities and the dilemma confronted by sensationalists centered on the problems of brute instinct and intelligence [Richards (1987), p. 22].

Sensationalists easily exploited the dilemma of the Cartesians, who wished to explain behavior on simple, natural principles and yet to capture the complexities it revealed [Richards (1987)], p. 24].

The sensationalists resolved the Cartesian dilemma by reformulating both physical and psychological theory. First, they admitted that animals and man were machines, though not composed of inert matter.

[...]

The sensationalists also introduced important epistemological and psychological reformations to the account of animal behavior. They argued that ideas were only copies of impressions received by sensory machines [Richards (1987), p. 25].

What is notable in these quoted words seems to be (1) that it is not crystal clear whether Richards understands the sensationalist as a sort of Cartesians, (2) even after all those heroic efforts, the problem of instincts was still unsolved in early 19th century, and (3) Richards formulated the issue as a dilemma between instinct and intelligence. (1) seems rather obvious, because if the sensationalists are never Cartesians, there would be no necessity for them to face the Cartesian dilemma. The most significant contribution of Richards to the history of the beast-machine controversy must be his interpretation of the sensationalists' influence on the 19th century evolution theory. It is simply impossible to delve into this huge problem area. Nevertheless, the following quote from Richards would be enough for grasping the big picture:

Evolutionists in the first half of the nineteenth century Jean-Baptiste de Lamarck, Pierre-Jean Cabanis, the young Charle Darwin, and Herbert Spencer synthesized mechanist and sensationalist notions by arguing that animals intelligently developed new habits, which through generations of practice gradually became innately determined and mechanically fixed instincts [Richards (1981), p. 201].



Finally, (3) is truly telling to my entire project. In Park (2011) and chapter 1 above, I tried to highlight the greatness of Avicenna and Peirce-Magnani in their treatments of estimation and abduction in animals as both instinct and intelligence at the same time. From that perspective, it is a false dilemma to formulate the issue as between instinct and intelligence. If so, the history of the problem of understanding the wonderful instincts of animals from Avicenna up until Aquinas (or even later times) was a history of the degeneration of the problem. Also, if such an interpretation is correct, then Peirce's and Magnani's notion of animal abduction can be viewed as a revival of the old Avicennian solution after a millennium. In order to present the whole story leading to Peirce-Magnani, of course, we need to go a long way. Not to mention the treatment of animal instincts in Darwin, other evolutionists in the 19th century, and neo-Darwinians, we should also cover Peirce-James controversy, the emergence of ethology, and many other big issues.

### Concluding Remarks

Let me conclude with a few words about what is truly remarkable in Peirce's and Magnani's rediscovery of the solution of the problem of wonderful instincts of animals. As we saw above, the example of young chicken was already found in Reimarus. Thus, what is significant in them cannot be the mere fact that they were introducing animals in the discussion of the existence of abductive instinct in humans. It should rather be found in their ingenuity of introducing abductive instinct in the context of debating about instinct. This last point seems to have some logical interest. But we should leave for another occasion.

**Acknowledgement.** I am indebted to the two anonymous reviewers for their useful comments. Also, I am grateful to Lorenzo Magnani's constructive criticisms.

### References

1. Avicenna, Rahman, F.: *Avicenna's Psychology. An English Translation of Kitāb Al-Najāt, Book II, Chapter VI with Historico-Philosophical Notes and Textual Improvements on the Cairo Edition.* Oxford University Press, London (1952)
2. Barad, J.: *Aquinas on the Nature and Treatment of Animals.* International Scholars Publications, San Francisco/London (1995)
3. Bayle, P.: *Historical and Critical Dictionary: Selections.* Translated, with an Introduction and Notes, Popkin, R. H. Hackett Publishing Company, Indianapolis (1991)
4. Black, D.: *Estimation (Wahm) in Avicenna: The Logical and Psychological Dimensions.* *Dialogue* 32, 219–258 (1993)
5. Black, D.: *Imagination and Estimation: Arabic Paradigms and Western Transformations.* *Topoi* 19, 59–75 (2000)
6. Clarke, D.M.: *Descartes's Theory of Mind.* Oxford University Press, Oxford (2003)

7. Cohen, L.D.: Descartes and Henry More on the Beast-Machine—A Translation of Their Correspondence Pertaining to Animal Automatism. *Annals of Science* 1, 48–61 (1936)
8. Cottingham, J.: A Brute to the Brutes: Descartes' Treatment of Animals. *Philosophy* 53, 551–559 (1978)
9. Descartes, R.: *The Philosophical Writings of Descartes*. Cottingham, J., Stoothoff, R., Murdoch, D. (trans.) vol. 2. Cambridge University Press, Cambridge (1985)
10. Diamond, S.: Four Hundred Years of Instinct Controversy. *Behavior Genetics* 4, 237–252 (1974)
11. Harrison, P.: Descartes on Animals. *The Philosophical Quarterly* 42, 219–227 (1992)
12. Magnani, L.: *Abduction, Reason, and Science: Processes of Discovery and Explanation*. Kluwer, New York (2001)
13. Magnani, L.: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. In: *Abductive Cognition*. Springer, Berlin (2009)
14. Magnani, L., Li, P. (eds.): *Model-Based Reasoning in Science, Technology and Medicine*. Springer, Berlin (2007)
15. Magnani, L., Carnielli, W., Pizzi, C. (eds.): *Model-Based Reasoning in Science and Technology*. SCI, vol. 314. Springer, Heidelberg (2010)
16. Malebranche
17. McCarthy, E.: *Medieval Light Theory and Optics and Duns Scotus' Treatment of Light in D. 13 of Book II of his Commentary on the Sentences*. Ph. D. dissertation, City University of New York (1976)
18. Paavola, S.: Peircean Abduction: Instinct or Inference? *Semiotica* 153(1/4), 131–154 (2005)
19. Park, W.: *Abduction and Estimation in Animals*, *Foundations of Science* (2011) (forthcoming), doi: 10.1007/s10699-011-9275-2 (published online, November 20, 2011)
20. Peirce, C.S.: *Collected Papers*, vols. 8. Hartshorne, C., Weiss P. vol. I-VI; Burks, A.W. vol. VII-VIII. Harvard University Press, Cambridge (1931-1958)
21. Radner, D., Radner, M.: *Animal Consciousness*. Prometheus Books, Amherst (1996)
22. Reimarus, H.S.: *Abhandlungen von den vornehmsten Wahrheiten der natürlichen Religion*, 5th edn. Frank und Schram, Tübingen (1754, 1782)
23. Reimarus, H.S.: *Allgemeine Betrachtunhen über die Triebe der Thiere*, 3rd edn. Bohn, Hamburg (1760, 1773)
24. Richards, R.J.: Influence of Sensationalist Tradition on Early Theories of the Evolution of Behavior. *Journal of the History of Ideas* 40, 85–105 (1979)
25. Richards, R.J.: Instinct and Intelligence in British Natural Theology: Some Contributions to Darwins Theory of the Evolution of Behavior. *Journal of the History of Biology* 14(2), 193–230 (1981)
26. Richards, R.J.: *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior*. The University of Chicago Press, Chicago (1987)
27. Rosenfield, L.C.: *From Beast-Machine to Man-Machine*, New and Enlarged Edition. Octagon Books, Inc., New York (1940, 1968)
28. Sepper, D.: Descartes and the Eclipse of Imagination, 1618-1630. *Journal of the History of Philosophy* 27, 379–403 (1989)
29. Sterrett, S.G.: Too Many Instincts: Contrasting Philosophical Views on Intelligence in Humans and Non-humans. *Journal of Experimental and Theoretical Artificial Intelligence* 14, 39–60 (2002)

# From Mindless Modeling to Scientific Models

## The Case of Emerging Models

Tommaso Bertolotti

**Abstract.** This paper aims at contributing to the ongoing epistemological debate on the nature of models by proposing an excursus from emerging to scientific modeling that will highlight the similarities between the two forms: this analysis will also allow to focus on the development of those traits that are instead typical of scientific models. The analysis of basic forms of modeling will show how even mindless processing of external reality does not provide passive descriptions but is rather a poetic aggression which constitutes external reality as the organism perceives it. In my argument I will make apparent how this poetic character is indeed common to both emerging and scientific modeling. My final contention will be that scientific endeavor, captured within Magnani's notion of *epistemic warfare*, is not characterized by a dramatic qualitative difference in the nature of models at play, but by a new conception and attitude displayed by scientists in their use of models, also coupled with the other fundamental scientific tool: the experiment.

## 1 Introduction

To provide an initial definition, we can agree that a model is something we use in order to gain some benefit in the understanding or explanation of something else, which we shall call the target. A model lets us understand the target, and behave consequently, in a way that would not be possible without it: different models usually optimize the understanding of different aspects of the target.

This definition of model should make it easy to appreciate how many situations we face everyday are tackled by making use of models: to deal with other people we make models of their minds and their intentions, to operate machinery we make models of their functioning, in the remote case of trying to escape from wildlife we

---

Tommaso Bertolotti

Department of Arts and Humanities, Philosophy Section  
and Computational Philosophy Laboratory, University of Pavia, Italy  
e-mail: [bertolotti@unipv.it](mailto:bertolotti@unipv.it)

make models of their hunting strategies and perceptual systems<sup>1</sup> to explore novel environments we make models of their spatial configurations, to mention only a few.

We make use of models in wide array of circumstances, but what all models actually share is a dimension of non-abstractness: we create them, or make use of models that were already constructed by other people, and models usually display a distributed nature, since they are either built on external, material supports (i.e. by means of artifacts, paper sheets, sound waves, body gestures) or, in case of mental models, encoded in brain wirings by synapses and chemicals (a mental map, for instance, is the mental simulation of the action of drawing a map – a powerful model construction activity – whose embodiment in the brain was made possible by the enhancement of human cognitive capabilities [Magnani, 2006]).

This is just as far as overt modeling is concerned: recent studies showed how the range of cognitive activities that can be classified as *model based cognition* is wider than expected, as many cases rely on forms of modeling that are not explicit to the agent's consciousness. Such use of tacit modeling is shared by animals as well, and is not a trait specific to human beings.

Conversely, a human-specific use of models seems to be displayed by scientific practice. In science, as we will see, models not only provide simplified descriptions of known phenomena, but often serve as an inferential tool to explore and constitute the target itself (as contended by [Magnani, 2012]). As we will see, current epistemology is engaged in fierce debates about *what* models are and *why* we can trust their empirical successfulness: I will suggest that an insightful approach to these questions can be derived from the analysis of what scientific models share with all other forms of modeling (of which they can be considered a peculiar subset), what their peculiarities are and, on a final note, whether these peculiarities can be acknowledged legitimately or should be rather thought of as an attitude of scientific endeavor towards models and the peculiar use it makes of them.

## 2 Models without Modelers?

“A model allows us to infer something about the thing modeled” is one of the most straightforward definitions of model available today [Holland, 1995, p. 33]<sup>2</sup>. The nature of scientific models has been one of the most debated topics in philosophy of science over the past few decades. A number of interesting solutions have been

<sup>1</sup> Movies offer many examples of this kind of modeling: fugitives sometimes cover their body in mud to prevent predators from targeting them because of their smell, other times they freeze to exploit the predator's – supposed – blindness to immobile objects: in all of these cases, in order to decide a course of action one has to construct a model of the predator he wants to avoid, considering relevant factors and factors that can be manipulated (for instance, a model suggesting that the predator can sense the prey's brainwaves is not put forward, as at the moment it would not allow to undertake any viable course of action).

<sup>2</sup> The simplicity of this definition must not sidetrack us: as I will contend in sections 3.2 and 3.3 together with [Morrison, 2009] and [Magnani, 2012], the model is the *conditio sine qua non* for poetically establishing a new scientific understandability which coincides with establishing the “borders” and the essence of the target-phenomenon itself.

put forward by different philosophers, but there seems to be a common issue: as if in a kind of name-fury, the common approach consists in branding the object of investigation with a new name, and consequently generate a new class of problems that do not relate specifically to scientific models but to their new “avatar”.<sup>3</sup>

An alternative suggestion might be to adopt a kind of bottom-up, naturalistic approach considering scientific models first of all simply as “models” (almost as if this was a primitive concept), and once their “behavior” is assessed to a satisfactory extent, analyze what makes them “scientific”.

The first bias that should be dispelled is the one characterizing the notion of model as associated with scientific modeling and thus intentional representation: models are often considered as the *intentional* output of high-level cognitive capacities, and their development requires the display of linguistic, mathematical and graphical abilities, plus a theoretical penchant towards explicit analogical reasoning and mental simulation, and a necessary ability to externalize and disembodify knowledge in the production of artifacts that serve as external representations. It must be acknowledged, though, that studies in distributed cognition already showed that such intentionality only describes part of the endeavor in scientific modeling: [Magnani, 2009] argued that the process of manipulative abductive modeling embodies a dimension of thinking through doing that is shared by certain mammals and birds too, as I will show in the second section of this paper.

Most models are considered to have a descriptive function of how (a particular) target system works, and serve the purpose of making successful prevision on future events based on causal relationships, whether they are held to be necessarily true accounts or merely effective “fictions”.

As suggested by Cartwright [Cartwright, 1983], scientific models can be understood as “prepared descriptions”. A model speaks to us in a different way than a non-actively modeled perception does: if compared to unprepared descriptions (that is, perception), our models seem to be wrong [Toon, 2010], yet the descriptions put forward by the model *prepare* for the application of mathematical structures, for instance. The actual spring I am playing with right now might be a fuller and more accurate phenomenon than what presented by the model, but the model can illuminate us about some traits that will be shared by *all* springs much better than the contemplation of the single, present spring.

---

<sup>3</sup> To say that scientific models are fictions leads us into examining the core problems of fiction [Woods, 2010; Woods and Rosales, 2010b; Contessa, 2010; Frigg, 2010a; Frigg, 2010b; Frigg, 2010c; Godfrey-Smith, 2009; Woods and Rosales, 2010a; Suárez, 2009; Suárez, 2010], to label them as representations opens the ancient issues of representation and mimesis [Chakravarty, 2010]: similar problems arise if we just apply to models classical definitions such as abstract entities [Giere, 1988; Giere, 2009; Giere, 2007] and idealizations [Portides, 2007; Weisberg, 2007; Mizrahi, 2009], to the more recent ones, as surrogates [Contessa, 2007], credible worlds [Sugden, 2009] [Kuorikoski and Lehtinen, 2009], missing systems [Mäki, 2009; Thomson-Jones, 2010], make-believe [Toon, 2010], parables [Cartwright, 2010], epistemic actions [Magnani, 2004a; Magnani, 2004b] or revealing capacities [Cartwright, 2009].

From a naturalistic perspective, the notion of model is intimately bound with the adoption of a future course of action. The model is valued to make predictions, and – in the case of emerging models<sup>4</sup> – it sometimes serves as a blueprint of the target system. The different regulation of a parameter in a model usually generates novel hypotheses, in an operation of eco-cognitive attunement between the external reality which is the target of the model and internal representation where the model is constructed, exploiting the common coding (in the sense of the expression received from [Chandrasekharan, 2009]) connecting the execution, perception and imagination of, for instance, motor impulses.

With this respect, some questions could be asked in order to contribute to the outgoing debate around the nature of models themselves: to what extent is it possible to produce models without the display of a conscious, intentional intelligence? Are human beings affected by this connotation of models? In this section I mean to show how a widespread biological feature such as the display of camouflage technique might suggest the emergence, even in organism poorly endowed at cognitive level, of actual models of how their predators' perceptual systems work.

As I will show in the third section, this suggests that the spontaneous construction of models can actually *emerge* at a mindless stage (as in perception) [Magnani, 2007a], and how some characteristics of such-conceived models do *a fortiori* apply to scientific modeling as well.

## 2.1 *Embodied Models of Agency Recognition: An Eco-Cognitive Necessity*

One of the greatest cognitive problems shared by all organism that are able to react (more or less plasticly) to external stimuli is external agency. A perceptual representation of one's external environment must as a matter of fact highlight the presence of external agents who might be predators, preys or competitors for available resources, in order to maximize one's chance of survival. To begin with, the notions of external environment and that of *other* agents are not absolute but rather they beg the "with respect to?" question. It is easy to understand that every agent can be part of any other agent's ecology: this is true even for us – human beings – since we are, as individuals, constituents of each other's environment.

Such *situated* and ecologically informed character of organisms' cognition is reverberated by the concept of affordance<sup>5</sup> [Gibson, 1979], which provides a

<sup>4</sup> In this paper I will refer to the emergence of models to indicate the spontaneous development of modeling representations by lower form of cognitions. My use of the term draws to the concept of emergence intended by [Holland, 1995; Holland, 1997]: it signifies those biological and interactional phenomena characterized by *constrained complexity*, in which an extremely articulated and complex resulting state is triggered by a limited number of simple components and rules.

<sup>5</sup> Originally belonging to the conceptual toolbox of ecological psychology, an affordance is a resource or chance that the environment presents to the "specific" organism, such as the availability of water or of finding recovery and concealment. Of course the same part of the environment offers different affordances to different organisms.

useful account of the role of the environment and external – also artifactual – objects and devices, as the source of action possibilities (constraints for allowable actions). Of course, different organisms apply to the kinds of local signs they can perceive a wide array of different modelings: while simpler organisms possess extremely simple models that encode an off-line representation of environmental affordances [Laurent, 2003], organisms with more plastic cognitive capacities and properly semiotic brains can produce symbolic models [Magnani, 2007a] or propositional-sentential models<sup>6</sup>.

The claim, that I will support in the following subsections, is that the construction of models emerges from low-level cognitive capacities, and can thus be thematized as an issue that is necessarily ecological in its nature, that is to say, it concerns the cognitive relationship of an organism and some aspects of its external reality. I will rely on contemporary studies concerning ethology and animal cognition to show that processes of agency recognition are basic forms of model-based cognition, and that a further proof of the emergence of these simple models is that they can be found as operationalized in (and therefore extracted from) camouflage mechanisms.

As I will soon prove, I think that an abductive framework is the fittest one to describe and investigate the formation of those models which animals produce, that emerge from signs they are able to recover from the environment. It seems therefore legitimate to speak of “animal abduction” [Magnani, 2007a].

## 2.2 *Emerging Animal Models as Abductive Representations*

Abduction, as understood within the Peircean framework, can be accounted for as the process of inferring certain facts and/or laws and hypotheses that render some sentences plausible, that explain (and also sometimes discover) some (eventually new) phenomenon or observation: it is the cognitive process in which hypotheses are formed and evaluated. Abductive cognition is active in many scientific disciplines but also in everyday reasoning: it is essential in scientific discovery, medical and non medical diagnosis, generation of causal explanations, generations of explanations for the behaviors of others, minds interplay, when for example we attribute intentions to others, empathy, analogy, emotions, as an appraisal of a given situation endowed with an explanatory or instrumental power, etc. The fact that the same abductive framework can be fruitful to investigate different levels of modeling seems to be another factor legitimizing the hope that the study of scientific models and the study of simpler mental models can fecundate each other and produce new insights.

In fact, abduction must not be regarded as a merely sentential inferential process: indeed, many studies explored the existence of *model-based* abductive processes, concerning the construction and the exploitation of internalized (or to the manipulation of external) models of diagrams, pictures and so on. These recent studies on abduction opened a much wider field of investigation concerning these multi-modal inferences: survival is an eco-cognitive task, requiring organisms to engage in a

---

<sup>6</sup> Our social cognition is for instance aided by models of behavior called moral templates [Magnani, 2007b]

relationship with the environment that is often a conflicting one – as I am claiming in this section – and the relationship with the environment is mediated by a series of cues the organism must make sense of in order to generate, even if tacitly, some knowledge it did not possess before [Magnani, 2009].

Traditionally, studies have concentrated on the human dimension of inferential reasoning, nevertheless Peirce himself had stressed several times how the concept of abduction was to be held relevant for a biologically wide description of cognition. *Making sense of signs* is an abductive activity that human beings share with any organism endowed with a nervous system or, on an even bigger perspective, any organism capable of reacting actively to modifications of its environment [7].

Animals rely on their senses alone in order to recognize the presence of other organisms in their surrounding. What senses pick up is not an immediate picture of external agency but a more or less rich complex of signs: these signs relate chiefly onto the senses of sight, smell and hearing (taste – when separated from smell – and touch appear to be cues more useful to proximally investigate the nature of an organism rather than to infer its presence: if I can taste it, it is probably already within reach of my tongue. Jacobson's organ, an apparatus for sensing pheromones and other chemicals compounds is yet another kind of sensorial organ).

The situation could be therefore described as follows: an animal must manage to detect the presence of other agents in order to maximize its own chance of survival, and such detection can only be inferred by operating upon meaningful signs. Signs are not associated randomly, but according to certain models that emerge in the animal system in connection with the stimuli it receives: they can be pre-wired or learnt. The following step consists in the operationalization of the correct affordances concerning the detected organism.

Millikan suggests that internal model-based representations in animal minds might mostly consist of PPR ("*Push-me pull-you*" representations), meaning they are both aimed at representing a state of affairs and at producing another, often suggesting a chance for behavior as received by the Gibsonian/affordance tradition [Millikan, 2004]. The indicative content of a PPR mental representation about external agent will therefore never be of the kind *Oh, look at that organism PERIOD* but rather *Look at that organism: should I attack/avoid/hurt/kill/eat it/mate with it?*:

An animal's action has to be initiated from the animal's own location. So in order to act, the animal has to take account of how the things to be acted on are related to itself, not just how they are related to one another. In the simplest cases, the relevant relation may consist merely in the affording situation's occurring *in roughly the same location and at the same time* as the animal's perception and consequent action. More typically, it will include a more specific relation to an affording object, such as a spatial relation, or a size relative to the animal's size, or a weight relative to the animal's weight or strength, and so forth [Millikan, 2004, p. 19].

<sup>7</sup> Even plants can be described as displaying a kind of embodied cognition [Calvo and Keijzer, 2009] and are therefore concerned by inferential causation as well. The perceptual and inferential horizon at play is of course radically incommunicable with respect to ours and to that of non-human animals we are able to refer to.



A striking connection between this kind of model and the “common coding” suggested by [Chandrasekharan, 2009], in her contention that:

[...] the execution, perception, and imagination of movements share a common representation (coding) in the brain. This coding leads to any one of these three (say perception of an external movement), automatically triggering the other two (imagination and execution of movement). One effect of this mechanism is that it allows any perceived external movement to be instantaneously replicated in body coordinates, generating a dynamic movement trace that can be used to generate an action response. The trace can also be used later for cognitive operations involving movement (action simulations). In this view, movement crosses the internal/external boundary as movement, and thus movement could be seen as a “lingua franca” that is shared across internal and external models, if both have movement components, as *they tend to do in science and engineering* [Chandrasekharan, 2009] p. 1061, italics not in the original text].

The way that animal modeling responds to the “common coding” criterion is clearly embryonic if compared to the use of models displayed by science and engineering: as for animal modeling, I would not go as far as claiming that “the trace can also be used later for cognitive operations involving movement (action simulations)”: this can be true for animals displaying more plastic cognitive abilities and learning mechanisms, individually or socially. A point in case, nevertheless, is the centrality of movement in both scientific and biological, emergent modeling. Movement is, as a matter of fact, at the core of the idea of manipulation, and therefore of experiment: manipulations impress movements on the external reality so that the resulting changes can work as “props” for the construction of models, since thinking through doing is often “thinking through moving”<sup>8</sup>, but movement is also the primary building block in emerging models, as the first difference to be discriminated is the difference between biological and non-biological movement.

As already contended, a trait that is typically displayed by biological emerging models is their *tenacity*. The success of artifacts such as fishing baits and hunting traps depends on the fact that most animals either display limited capacities for learning and revising their inner models, or are not able to share their advancements

---

<sup>8</sup> [Magnani, 2009] stresses the centrality of manipulative abduction and the problem of thinking through doing in the scientific enterprise. The role of manipulation and thinking through doing is crucial also in the expression of the most advanced kinds of models displayed by animals: corvids, for instance, do not only exhibit exceptional ability in the creative use of tools, making smart use of non-natural items (e.g. aluminum strips) [Weir and Kacelnik, 2006], but can also operationalize complex mental representations such as Archimedes’ principle: [Bird and Emery, 2009] show how rooks can drop stones in a container filled with water so to raise the water level and attain a floating prey. They are also aware that larger stones cause a higher raise in water level than small ones. This seems to be a more sophisticated model-based activity, because if rooks are able to operationalize a model corresponding to Archimedes’ principle, then we should concede that they *possess* that model: they cannot relate to the model in theoretical-sentential way typical of humans beings, but it is nevertheless encoded in neural systems and accessed in instances of thinking through *through doing*.

with their conspecifics.<sup>9</sup> Emerging models display such tenacity, and change over long period of times, inasmuch they are favored by natural selection if they work most of the time (that is, if they do not cause the systematic death of the organisms who entertain certain models).

As far as the biological and pre-linguistic levels are concerned, it can be argued that those emerging models do not matter for their *truth-reliability* but rather for their *fitness-reliability*.<sup>10</sup> [Sage, 2004]. From a biological outlook (which is often engaged by human beings as well) the favored inference is the most successful inference, the one leading to survival. For instance, not noticing the presence of a predator, not entertaining any form of PPR representation concerning it – and thus not reacting – might be the best way to avoid being noticed in turn and killed: in this case, the potential prey's proto-belief is clearly false, and yet successful. Similarly,

[...] cautious cognitive faculty that “over detects” dangerous predators (frequently generating the false belief that a predator is nearby) may generate an abundance of false beliefs, though it may turn out to be adaptive because these false beliefs increase an organism's inclusive fitness (p. 97). [...] The abundance of adaptive false beliefs gives us reason to doubt that true beliefs are more likely to increase an organism's inclusive fitness than are false beliefs (p. 102) [Sage, 2004].

It is interesting to note how abduction, as it is not a truth-preserving inference, perfectly depicts such inferential scenarios: considering as premises beliefs held as true by the subjects, abduction generates emerging models which may not be true and yet be endowed with a powerful *fitness-reliability* for the organism's welfare. The same happens in human reasoning: in peculiar settings we may produce models of a target without having gathered necessary evidence, as in the case of hasty generalizations [Woods, 2004]. Such hastily generalized models (concerning for instance generalizations about women who are not able to drive through traffic, or concerning big felines that are afraid of water, or generalized models of bombs that can be defused by cutting the yellow cable – no wait, was it the blue one?) can be valuable for their contribution to the agent's fitness inasmuch they can help her make a decision that saves her life, but do not benefit the *epistemic welfare* of the agent herself.

<sup>9</sup> Conversely, some species (typically rodents and birds which share their habitats with human beings) are said to be endowed with a kind of “culture”, inasmuch as they show a clear predisposition towards constructing models that are actively tuned with ecological necessities and sharing them with conspecifics by means of social learning, observation etc. [Heyes, 1993].

<sup>10</sup> Especially when comparing animal fitness and cultural evolution I suggest that the concept should be understood in a *loosely Darwinian* connotation. Besides, when considering animal fitness, a local, mid-term conception could be acceptable as well with the respect of this investigation: i.e. assimilating the concept of *fitness* to the one of *welfare*. In the case of science, truth-reliability could be said to go together with a kind of meaning-reliability, inasmuch as models are valued also according to epistemic meters such as coherence, consistency etc.

### 2.3 *Emerging Models: Useful Instruments or Fictions?*

At this point of my analysis of emerging models, a proper demarcation between what we just described and scientific modeling should be stressed. If we transpose into the domain of scientific models the importance of *fitness-reliability*, one could think that it is indeed possible – to paraphrase a verse of Shakespeare’s that is indeed much loved in the debate on the nature of models – to catch a “carp of truth” with a “bait of falsehood”. On the one hand, as contended by Giere [Giere, 2009], the instrumental use of some models that are known to be fictional does not entail that the whole system should be labeled as fiction. Some models explicitly aim at simplifying calculations by offering a different systems to refer to:

Applying the method of image charges, one replaces the original model with a model in which the infinite metal plate is replaced by a “fictional” negative charge placed symmetrically on the other side of where the surface had been in the original model. The solution to the problem using the new model, in full accord with electrostatic theory, is exactly the same as if one had solved the mathematically more difficult problem using the initially suggested model. What is meant by calling the negative charge in the second model “fictional”? As a component of a model, the image charge in the second model is no more and no less fictional than the positive point charge and infinite metal surface in the original model. It is telling that textbooks do not refer to the latter as fictional although they are clearly physically impossible entities. My analysis of the situation is that the original model is understood to be an idealized representation of a concrete system. The concrete system would only have counterparts to the original positive charge and conducting surface. Relative to this suggested concrete system, the negative charge in the second model is called “fictional” because it would have no counterpart in the assumed concrete system. On this understanding of the situation, there is no basis for calling either model as a whole a work of fiction [Giere, 2009].

If we compare the solution of a problem (maybe even the same problem, say a physics test) resorting to a hasty generalization (“My teacher always follows the same pattern in distributing the right answer in quizzes: the right answer must be the third!”) or by using the system described by Giere, we can easily say which one of the two is closer to the idea of fiction. The first model just provides a (highly fallacious) *fitness-reliable* way to plug an ignorance leak with no commitment towards the relevant target system, while the second model projects a scientific and *truth-reliable* structure of understandability on the target system in order to make the problem more easily solvable.

If we go back to the origin of the demarcation, it could be said that I drew an analogy between animals’ *fitness-reliable* models and hasty generalizations, but it would be actually unfair and wrong to depreciate the dignity of emerging animal models by comparing them to hasty generalizations: coherently with Millikan’s observations, I argue that animals’ emerging models concerning external agency strictly depend on the *extreme situatedness* of their cognitive capacities. Most emerging models cannot therefore be separated from the *here and now* relationship between the organism who entertains the model and the target of the model (that is, its predators or preys and the way they relate to the cognizing agent). With this respect, it could be suggested that models, even from a nearly mindless perspective, can be

seen in turn be seen as a “model” of the target system’s affordances. As maintained by [Laurent, 2003] and coherently with what I have suggested so far, an emerging model can be conceived as simulation of the target’s affordances, thus operationalized to allow the prediction of future events connecting to those very affordances. For instance, a model could represent a property of the target such as “being able to detect only moving agents”, by the related affordance “escape from the predator by remaining immobile”, and direct the behavior of the cognizing agent. Analogously with Chandrasekharan’s use of the concept of “common coding”, such abductive representations appear in fact to be the product of situated abductive inferences, peculiar inasmuch “they tell in *one undifferentiated breath* both what the case is and what to do about it” and they “represent the relation of the representing animal itself to whatever else they also represent” [Millikan, 2004] p. 20, italics not in the original text]. This kind of inferential process, residing in the coupling of the detector and the detected, is not based upon a random appraisal of an animal’s semiotic cloud, but specific sign configurations match certain affordances, which ultimately trace back to the desired property. Jacob and Jannerod’s description seems particularly illuminating:

Property *G* matters to the survival of the animal (e.g. a sexually active male competitor or an insect to capture). The animal’s sensory mechanism, however, responds to instantiations of property *F*, not property *G*. Often enough in the animal’s ecology, instantiations of *F* coincide with instantiations of *G*. So detecting an *F* is a good cue if what enhances the animal’s fitness is to produce a behavioral response in the presence of a *G* [Jacob and Jeannerod, 2003] p. 8].

The hypothesis about the presence of an agent who detains the property *G* is *abduced* on the basis of one or more *perceptible* properties *F* that usually signify the relevant properties according to the emerging model. If an organism is hunted as a prey or avoided as a predator because of a property *G*, it must try to reduce the occurrences of the properties signaling their characteristic, and this varies widely from organism to organism.

The emerging model prescribes that something is taken to be symptomatic of something else on the base of some regularity, and this *cognition* can be instinctual or plastically shaped by learning.

Before moving on to the next subsection and witness and analyze the role of emerging models in camouflage mechanisms, we should take advantage of the last considerations about emerging models to reflect about the impossible ontological abstractness of models: as I showed, emerging models are necessarily situated and have often a clear operative and heuristic role (they guide the cognizant’s behavior) while we cannot say much about their representational scope (it would be rather speculative to say something like “The sepia represents anything moving as a possible prey”: this is about *our* model of the sepia’s cognitive capacities, and not about the actual sepia’s model). Our simple models, on the other hand, do have an abstracting quality: a simple map, or a model of a person’s way of reacting to some news, can be accessed and constructed independently from the proximity and the actuality of the target, but they are nevertheless *connected* with their target, produced by a *peculiar* cognition and externalized by a modification of some support (be it neural,

physical, interactional etc.): in brief, from emerging models we can already learn that models are *necessarily* someone's models of something.

## 2.4 *Camouflage as the Strategic Use of Models in Nature*

The discussion on natural models that I began to lay out in subsection 2.2 seems to legitimize the claim that every agent has a twofold inferential relevance, active and passive: on the one hand, it disperses signs *out* in its environment, on the other hand it receives and processes signs *from* other organisms: these signs are received and used by model-based agency recognition cognitive processes. The first mechanism must be minimized (spreading out signs) while the second (recovering signs) maximized either to counteract predation either to avoid being spotted by a potential prey. If by now we exclude that model-building is an activity engaged by human beings alone, we can concur in saying that every organisms (more or less explicitly) attempts at producing valuable models of other organisms' behaviors.<sup>11</sup>

If we assume that an organism is endowed with – mostly model based – abductive cognitive systems aimed at the detection and identification of other agents in its surroundings, we can suppose that these systems operate within a determinate threshold selecting stimuli which activate their inferential processing. Signs that fall within these abductive thresholds are likely to produce in the organism an internal representation involving some kind of awareness about a particular nearby agent, which emerges as a model of that agency.

In order to maximize their chances of not being discovered by agency recognition systems, certain organisms were favored by natural selection into modulating their semiotic footstep and let out signs that can be few and deceiving (falling under the inferential threshold of other agents, so that they do not trigger any positive agency-detection response) or meant to overwhelm and saturate the agent's abductive threshold. This is about a strategic use of models: the peculiar display of signs is meant to hinder the application of an unsuitable model, or block the development of any model at all.

Stevens and Merilaita define camouflage as “all strategies involved in concealment, including prevention of detection and recognition” [Stevens and Merilaita, 2009, p. 424]: they maintain that camouflage should be analyzed with respect to its *function* and *mechanism*, thereby stressing the

<sup>11</sup> When dealing with animal cognitive faculties, especially if rather speculatively, descriptions may always be tainted by the typical anthropomorphism of the observers “psychological” explanations. Unfortunately, even when interested in animal cognition, researchers fall necessarily victims of an uncontrolled, “biocentric” anthropomorphism: there is always the risk of attributing to animals (and of course infants) our own human concepts and thus misunderstanding their specific cognitive skills [Rivas and Burghardt, 2002]. In my analysis of camouflage, when I fall in anthropomorphic slips, I assume it is because of the lack of better words (e.g. when I say that an organism “attempts” at something, I mean to indicate a telic action rather than an explicit volition): it is an instrumental anthropomorphism and not a psychological, gnoseological or ontological one.

relevance of local semiotic interactions<sup>12</sup>. As a consequence, it can be said most dynamics broadly labeled as camouflage can be aimed to prevent detection, avoid recognition or to avert the opponent from operationalizing a PPR representation, that is correctly exploiting the affordances expressed by that particular model.<sup>13</sup>

*Crypsis* usually individuates those processes in which the initial attempt is to prevent detection. When we intuitively think of camouflage, we usually think of *crypsis*. If we mean to describe *crypsis* in a semiotic-abductive framework, we could say it functions by downplaying signs so that they not activate other organisms' agency detectors: those signs do not nudge the cognitive system into abducting their origin, and therefore do not trigger the production of a PPR representation that could prove lethal for the camouflaged organism, or alert the prey if *crypsis* is enacted by a predator. *Crypsis* is very effective inasmuch as it is intuitively hard to develop an emerging model upon something which is simply not there: no prop is provided at all for the construction of the model.

*Masquerade* is a semiotically different kind of camouflage, inasmuch organisms do not attempt to merge with the background: conversely, they provide into the environment signs that make them easily detectable, but "their bearers are misidentified as either inedible objects by their predators, or as innocuous objects by their prey" [Skelhorn *et al.*, 2010, p. 1]: they provide, that is, patent props to operationalize recognition models that are actually misleading.

Forms of *kinesthetic camouflage* also exist, and they rely on the alteration of a particular subpart of the organism's semiotic shadow: their aim is not to prevent an organism from being detected nor to be recognized, but to prevent an effective prediction of their spatial bearings [Srinivasan and Davey, 1995]. "Motion camouflage is a strategy whereby an aggressor moves towards a target while appearing stationary to the target except for the inevitable change in perceived size of the aggressor as it approaches" [Glendinning, 2004, p. 477].

If *crypsis* produces signs that are not configured as cues for possible abductions, *masquerade* tactics offer indeed a profusion of signs likely to be picked up by other agents, but that are not to be processed as relevant by agency recognition but are instead actively acknowledged as inert objects belonging to the environment. What is at stake is not the possibility of performing abductions upon a configuration (or non-configuration) of signs in the local environment, but the quality of such abductive inference, and the reliability of the consequent PPR representation relating to the embodied model of agency recognition. A certain *counter-factuality* could be ascribed to the kinds of PPR representation (or the lack thereof, i.e. when a predator or prey is not spotted) triggered by camouflage, insofar as they either depict organisms differently from their real nature or they fail to depict them at all, when they are present.

The strategic model-based warfare incorporated by camouflage dynamics concludes when the model employed or constructed by one of the organisms fails in the

<sup>12</sup> "In defining different forms of camouflage, we use the term 'function' to describe broadly what the adaptation may do (e.g. breaking up form, distracting attention), and the term 'mechanism' to refer to specific perceptual processes (e.g. exploiting edge detection mechanisms, lateral inhibition)" [Stevens and Merilaita, 2009, p. 424].

<sup>13</sup> Coherently with the claim expressed by [Laurent, 2003], quoted in subsection 2.3.

representation of the target, thus instancing an occurrence of what could be seen as “negation as death”: negation of the model because of the death of the modeler.<sup>14</sup>

To put it another way, it can be said that a successfully camouflaged agent managed rearranged its imprint of signs according to the model of agency-recognition of its prey/predator: its predators and preys, in turn, were not able to spot it (or failed to recognize it) because their model of agency-recognition was not in tune with the target. But if we consider this kind of models in their relationship between the representation and the target system, can it be said that they display a fictional nature, even when they contribute to producing clearly counterfactual beliefs? I would not say so: contemporary philosophical tradition – in particular phenomenology – has stressed how, despite the fact that perception can trick us and become misperception, it can never be said to be *wrong*. If perception cannot be wrong, then it cannot be fictional either, inasmuch as the fictionality of a model can be – if ever – assessed necessarily *a posteriori*.

### 3 The Naturalness of Scientific Models

In the previous section I showed how the presence of emerging models in basic animal cognitive faculties can be inferred from the analysis of eco-cognitive mechanisms such as camouflage interactions. In this section I intend to go to the heart of the matter and generalize to human modeling part of the scenario we have described so far, highlighting the continuum between emerging modeling and scientific models. In subsection 3.1 I will contend that the relevance of models in many cognitive tasks makes unviable their reduction to fictions: this applies, *a fortiori*, to scientific models. Subsection 3.2 briefly analyzes how both emerging and scientific modeling is the preparing step for mathematical abstraction.<sup>15</sup> In the last subsection, 3.3, I will deal with the peculiar nature of scientific models analyzing their origins in the scientific revolution, contending that their specificity relies not in the models *per se* but in the attitude towards them in scientific practice.

#### 3.1 *All Human Knowledge Is a (Sometimes) Virtuous Distortion (and a Model Too)*

In [Bertolotti and Magnani, 2010] we provided an extended excursus on the building blocks of inferential model-based reasoning, that is on how human beings process raw information in their environment according to model-based cognitive

<sup>14</sup> [Magnani, 2012], in this book, examines the relationship between fictionality and falsehood in terms of the logical concept of “negation as failure” [Clark, 1978; Magnani, 2001]. Negation as failure is a *weak* form of falsification, compared for instance to a counterexample that would negate the model in its matter: the failure of the model means that it just stops working, even if it was not *proven* wrong: emerging models admit another kind of (paradoxically) *weak* negation, that is “negation as death” of the cognizing agent.

<sup>15</sup> As meant by [Morrison, 2009]: a projection of meaning which reconfigures the phenomena, as opposed to models by idealization of the target.

heuristics that result in the complex cognitive artifact that is our perception: this should make us consider how, even such models that could be called emerging models in animals – to stress their instinctual and immediate nature – are shared by human beings as well. As suggested by [Woods and Rosales, 2010b], there is a continuum connecting basic mental models and the most complex scientific models – and they all share the dimension of being a virtuous distortion – so that the analysis of the latter benefits the former and vice versa. Even perceptual knowledge for human beings (though at a very basic level, before being influenced by higher structures such as symbols, languages, theories etc.) is a model based activity, whose data processing rate “is in the neighborhood of 11 million bits per second. For any of those seconds, something fewer than 40 bits make their way into consciousness. Consciousness therefore is highly entropic, a thermodynamically costly state for a human system to be in. At any given time, there is an extraordinary quantity of information processed by the human system, which consciousness cannot gain access to”, and the result is that, fundamentally, knowledge is not only a product of “information-processing” but also an “information suppressor” [p. 17].

It is a model-based activity as far as the picture accessed by consciousness is both an abstraction (inasmuch some true stimuli are suppressed) and an idealization (since some fake stimuli are introduced, as in optic illusions): this phenomenon is referred to in cognitive science as semi-encapsulation of information [Raftopoulos, 2001b; Raftopoulos, 2001a; Albertazzi *et al.*, 2011].

In this case, if we compare perceptual models with scientific models, the former are clearly not the descriptions of missing systems, even if “competent practitioners” could assure us that albeit perception being an apparently “accurate description of an actual, concrete system (or kind of system) from the domain of inquiry, [...] there are no actual, concrete systems in the world around us fitting the description it contains” [Thomson-Jones, 2010, p. 283]. Perception could match this description, but such a categorization falls short of any value if the system is the only available one, produced in order to make sense of the world, and operationalized *as if* it coincided with the target system. Other animals embody different perceptual models than ours. For instance, mosquitos are endowed with a heat-seeking sensorial apparatus, while bats are known for their ability to perceive ultrasounds: clearly their models of external reality differ from ours, but this should not lead us into defining every perceptual model as a description of a missing system (albeit the definition would fit, it would be of little profit).

Clearly, with respect to scientific models, the kind of models we just described is affected by two fundamental properties: dominant transparency and a consequent lack of plasticity. The models of perception enacted by our cognitive systems are not negotiable: we cannot affect our perception naturally, but by means of technological artifacts it is possible to integrate out embodied perceptual models, for instance by use of heat-sensing (infrared) goggles.

Notwithstanding the existence of differences between individuals, perception considered as an unintentional model is usually *de facto* shared by all organisms belonging to the same species, but lacks a characteristic that is typical of scientific models: the latter are in fact basically intentionally shared (even if characterized by



embodied cognitive acts transparent to consciousness), and their plasticity derives from the fact of being arguable, revisable and withdrawable, just as the inferences that underpin them.

Woods and Rosales seem to have pinned the issue in a very clear way:

Here, then, is the basic picture. Knowledge is the fruit of information-processing. But it is also an information suppressor. There is a basic reason for this. Consider the particular case of conscious representational knowledge. If what is suppressed by our cognitive processes were admitted to consciousness and placed in the relevant representational state, it would overload awareness and crash the representation [...] This supports the abstraction thesis: *A cognitive state is an abstraction from an information state.* There is another way of saying the same thing: *A cognitive state is a model of an informational environment.* [Woods and Rosales, 2010b] p. 17]

Our knowledge-gathering sensorial experience emerged naturally as an information-processing activity preventing *overloads* that would just *crash* the representation. No more, no less. And, since we are not prone to give up on giving our assent to our *now appearing heavily mediated* percepts, we should – out of epistemic honesty – refrain from refusing scientific models the basic, unquestioned, acceptance we accept perception with, because the underlying structure is the same (perchance better controlled and executed in scientific models).

What we now propose is that scientific models mimic the structures and processes of cognition quite generally, and that a fully worked up model-based scientific theory would capture with some precision the constructive impossibility of knowing a thing through an awareness of most of what's true of it. With perception again as our guide, knowing of the world what you do when you see the robin in the tree is, in comparative terms, knowing hardly anything that's true of it. Such knowledge – a conscious awareness of the disclosures of your five senses – beckons paradoxically. *It supplies you with ample knowledge on practically no information.* [...] In these respects, the abstractions of model-based science mimic the abstractive character of knowledge in general. *If, as we ourselves suppose, the abstractive character of perceptual states doesn't deny them cognitive value, why would the abstractive character of model-based reasoning deny it cognitive value? After all, shouldn't it be that what's sauce for the goose is sauce for the gander?* [Woods and Rosales, 2010b] p. 18–19, italics not in the original text]

Woods and Rosales' argument is philosophically powerful: they do not utterly deny that there might be some kind of a fictional nature in model-based reasoning, but to assume the fictional dimension as its *characterizing* trait would imply to drastically and methodically question the trust we place in something as basic as perception and every specimen of knowledge. This is a clearly feasible philosophical endeavor which has often been practiced in the history of philosophy, but its value in questioning a process that science has *proved* to be effective for the past four centuries is a whole different kettle of fish.

Of course, their (and my) claim is not that there are *no* fictional models whatsoever. First of all, as explained by Giere [Giere, 2009], we already have a splendid and non-ambiguous word to define fictional models, that are conceived and shared

as such: *fiction*. A novel, classified in libraries and bookstores under the label “fiction”, is a *work of modeling* that clearly creates a universe of meaning, goes even as far as to *constitute* a whole new series of phenomena (think of Tolkien’s saga, for instance), but the author’s intention was never to describe an actual part of external reality, instead of science. Clearly, as contended by [Giere, 2009] and by [Magnani, 2012], in this book, the purpose of consciously fictional modeling is different from that of scientific – and perceptual – modeling. In the latter case, the aim is to construct models of reality that *directly mediate* the course of our understanding and of our action about that *particular* reality: conversely, in the former case – that is, fictional modeling – there might be indeed an attempt to *refer* to an existing external reality, for instance adopting a moral framework, but this is not the main purpose of the fictional model and, at its best, it is not straightforward but *inferentially rich* if not openly *ambiguous*.

This matter is worth a small digression: if we take as an example Watson and Crick’s double-helix model of the DNA, we know was constructed in order to explain a certain phenomenon – the base-pairing in DNA molecules.<sup>16</sup> It is an abstraction (as understood by [Morrison, 2009]), because the model is not achieved by analyzing the external phenomenon and subtracting local constraints to achieve a functional description of the general case (e.g. the ideal pendulum, ideal spring etc.), but the model is what instantiates the fact that the phenomenon can be understood and subject to theory *as such*: to say it another way, it is the model that traces the theoretical borderlines of its target, which is in turn defined *through* the model, and subsequently proved to be true (as successful) by its engagement in scientific practice.

In this sense, a scientific model creates a phenomenon by *abductively* configuring it out of external reality (but, as we saw, is not this what perception does?) and then suggests how to look for confirmations of the model itself by investigating the consistency of phenomenon, now isolated and defined from reality.<sup>17</sup>

Let us compare it to something which is a model *and* a work of fiction at the same time: “Animal farm” by George Orwell. The book is intended as a satirical novel, and focuses on a modeled society ruled by anthropomorphic animals. If we analyze

<sup>16</sup> The invention of the particular model is just a stage within the *epistemic warfare* that [Magnani, 2012] introduces in this book. It is not invented *ex nihilo* and found to be matching with the desired phenomenon, but it is produced by a continuous series of attempts, successes, failures that cause the revision of precedent models and lucky events that serve as a “prop” to invent the new model (a snake apparently gave Watson & Crick the intuition about the shape of the DNA molecule!).

<sup>17</sup> A fundamental problem of epistemology was nonchalantly touched here, heavily connected to the issue of the so-called theory-ladenness of scientific facts (contended by Hanson, Popper, Lakatos and Kuhn among many others), and partially explains why scientific truth must be considered as provisional: if a model entails at least partially then phenomena it refers to, then it is somewhat self-supporting. Not in a vicious, question begging fashion, but because the model configures the range of phenomena that are *readable* to serve as falsification or corroboration of the model itself. Models are found to be outdated when new facts emerge to show it was wrong, usually from new observations (because of technological improvements) or from “blind-spots” of the model itself.

the structure of the model in a strict way, it is a fictional model of a fictive target, that is the farm ruled by anthropomorphic animals. To echo Giere's analysis, if we limited ourselves to diss the novel basing on the fact that there is no such known thing as a farm ruled by animals, all we would display is poor understanding of the word "fiction".

Furthermore, one could argue that the model has an allegorical meaning, and that the target system it refers to is indeed human society: three observations should be made to this respect. First, the model would indeed refer to some *moral* characteristics of human societies and not to some exact traits, and it would display a descriptive role rather than the predictive/heuristic role that models play in science; second, even the process of inferring some facts from the model onto external reality is, as we said, *inferentially rich* – that is, according to one's personal sensibility a more or less grim picture of human society may emerge (or even an absolutely positive or physiological one!). Last, but not least, the correct pathway to be taken to appreciate the model is not unique: usually depending on personal predispositions and culture, one can appreciate the book as a fictional model of a fictive system (like children usually do), or as a fictional model allegorical relating to an existing target<sup>18</sup>

The mechanisms we just briefly described do not apply to scientific modeling: what the model refers to is unambiguous and most of the time there are no concurrent ways of exploiting the same model to say different things about one target system, and it is hard to swing a model as it is between different targets and sometimes accept its applicability and sometimes not. This is why, considering the premises we have been following so far, Cartwright's claim about models being like parables is rather thought-provoking:

[...] in many cases the correct lessons to be drawn may be more abstract than those described immediately in the concrete situation of the model. But seldom can we really cast the models as fables because the moral is not written in. They are rather like parables, where the prescription for drawing the right lesson must come from elsewhere. Theory can help here, as can a wealth of other cases to look to, and having a good set of well-understood more abstract concepts to hand will play a big role. So the good news that one can move from falsehood in a model to truth by climbing up the ladder of abstraction is considerably dampened by the fact that the model generally does not tell us which ladder of abstraction to use and how far to climb [Cartwright, 2010].

As Cartwright suggests in the paper, parables indeed differ from fables inasmuch as fables usually end with an explicitly stated moral, that is more or less entailed by the development of the fable. I would say that the similarity between parables and models stops here, or rather, could be expressed as "if models are fictions they are not fictional as a fable is". Conversely, the idea according to which there is a "right lesson" – and therefore a "wrong lesson" – to be drawn from a model is interesting but must be compared with a correct *dynamic* understanding of scientific endeavor.

<sup>18</sup> This is the key to the success of recent computer animation major pictures among public of different age: children appreciate them as 100% fiction, while grow-ups can enjoy both the fictive nature, and the extent to which they mirror existing systems.

As I have already suggested, scientific enterprise is rightly identified by Mag-nani, in this book, as a state of *epistemic warfare* which sees scientists engaged in an aggressive battle *against* nature: the idea of warfare nicely captures the dynamic dimension of scientific endeavor, aimed at producing valuable knowledge about the various fields of investigation. This specificity of science can be traced back to the famous quotation by Francis Bacon about the “vexation” of Nature by the scientists themselves: [Pesic, 1990] argues in fact that Bacon’s ideal was not of science violently aggressing an unarmed victim, but rather of a “heroic mutual struggle” (p. 81). Models are some of the most used and useful weapons of this struggle. As in any state of warfare, it can sometimes be the case to choose a preexistent weapon (i.e. model), while some circumstances might require the development of completely new weapons (i.e. models).

If we consider Cartwright’s conclusion in this perspective, something seems to be slightly puzzling: “Theory can help here,” she said “as can a wealth of other cases to look to, and having a good set of well-understood more abstract concepts to hand will play a big role.” According to the line of thought we followed so far, this claim is puzzling. Applying a charity principle, I suggest that – if we consider the dynamic nature of science – Cartwright’s claim is actually a self-evident truth. It goes without saying that theory helps in the selection and the construction of a good model (with the addiction of all other more or less accidental factors), and it can also happen that according to their level of expertise two scientists can make different sense of the same model (usually with the help of additional manipulation). What I contend is that the production of a model, which in turn as I stressed little earlier produces much of the target phenomenon itself, cannot be separated from the act of interpretation of the model. To draw the “right lesson” from a model is just another way of saying one developed a successful model, while drawing the “wrong lesson” means that one developed and applied an unfruitful model, which did not provide any reliable understanding of the target (nor configured the target as properly understandable). What I find utterly puzzling is that such a distinction would make sense in a rather unrealistic static conception of science, where the modeler and who makes use of the model are not the same person– nor they belong to the same party like the laboratory group – as it happens in parables!

If we consider Jesus’ preaching in a pragmatic-historical framework, it can be easily seen that Jesus did not admit a good lesson *and* a bad lesson to be drawn from his parables, and at least in one occasion he rebuked his disciples several when they would not understand the meaning of a parable.<sup>19</sup> Today, as we cannot ask Jesus to explain parables to us, they are sometimes straightforward, sometimes inferentially rich, and some other times they are ambiguous *tout court*: different interpretations of famous parables such as the one of the *workers in the vine* or that of the *buried talents* played a role in opposing different Christian confessions over the centuries.

Coherently, we can say that the strength of a parable resides partly in their being inferentially rich (they have been able to tell something new to Christians spreading over five continents and twenty centuries), while this is not necessarily a quality in

<sup>19</sup> As it happens in Mark 15:15-16, “Peter said, ‘Explain the parable to us.’ ‘Are you still so dull?’ Jesus asked them.”

scientific models, whose desired qualities concern more the possibility of individualizing capacities, so that models can guarantee fruitful predictions. The discovery of capacities is ultimately linked to the development of models according to Cartwright herself [Cartwright, 2009], and is echoed by Sugden: a “[...] satisfactory isolation, then, allows a real relationship of cause and effect to be demonstrated in an environment in which this relationship is stable. In more natural conditions, this relationship is only a latent capacity which may be switched on or off by other factors; but the capacity itself is stable across a range of possible circumstances. Thus, the model provides a ‘theoretical grounding’ for a general hypothesis about the world” [Sugden, 2009, p. 20]. I think I managed to explain why parables and other kinds of consciously fictional accounts of real or fictive targets do not help to isolate any capacity.

### 3.2 *Both Emerging Models and Scientific Models Prepare for Mathematical Abstraction*

A fundamental trait of contemporary scientific modeling, as stressed by [Morrison, 2009], is their being a support for mathematical abstraction: albeit neopythagorean intuitions possess an unmistakable philosophically romantic connotation, the mathematization of perception is necessarily mediated by a modeling structure and cannot be *naturally* given. As I will contend, the fact that even simple percepts often offer a significant mathematical meaning is a sign of how both emerging and scientific models are what supports the *creation* of meaning, for instance by mathematical abstraction.

The origin and status of numbers is indeed a fundamental problem of philosophy and philosophy of mathematics, but it will suffice for this analysis to agree with [Holland, 1995] in his claim that numbering is one of the most basic examples of emerging models: numbers emerge from a model of external reality that affords the isolation of quantities and the abstracting step that lets the cognizant grasp that quantities are the same even if the actual objects are different.

As proven by recent cognitive research, organisms’ basic modeling capabilities (that already offer a what [Cartwright, 1983] would call “prepared descriptions”) afford more elaborate inferential processes, in spite of their being situated at a low cognitive level [Dehaene, 1997; Dehaene *et al.*, 1999]: as suggested by [De Cruz, 2006, p. 157], “the human capacity for mathematics is a category-specific domain of knowledge, hard-wired in the brain, which can be explained as the result of natural selection”. Mathematical modeling could therefore be seen as a step in the evolution of human cognition, which had risen before the full development of conscience as we know it (considered as a necessary condition for scientific endeavor). Significant research was recently lead on a phenomenon called “subitization” [Davis and Holmes, 2005]: it relates to the numerical estimations that our cognitive systems can perform without actual counting. Usually, human beings are able to recognize by subitization quantities that amount up to four units. In a loosely

Pythagorean speculation, this kind of phenomenon could be understood as a tacit modeling connecting even and uneven quantities to agency detection:

Crucially, most biologically important objects, such as predator or prey, are symmetrical and, in this respect, sensitivity to symmetry may have evolved because it is crucial for discriminating living organisms from inanimate objects. In fact, symmetry seems to act as an early warning system that directs the visual system to further scrutinize an object until full recognition has occurred. Mirror symmetry is thought to have special status in human perception, precisely because it is such an important cue as to the presence of natural organisms [Hodgson, 2009, p. 94].

Subitization mechanisms were subsequently hybridized with other kinds of externalized modeling, either very complex or as simple as using a pencil to count a line of dots dividing it into groups of three or four units [Kirsh, 1995].

As I already suggested and as I will further stress in the final subsection, we can agree with Cartwright in seeing models as an intentional, emerging or hybrid “prepared description” of the target which lets us perform inferential activities about new features of the target.<sup>20</sup> [Morrison, 2009] nicely extends this insight by showing how this cognitive preparation of a mediating structure to understand the target is fundamentally creative in those cases calling for the mathematization of the phenomenon.

In situations like this where we have mathematical abstractions that are *necessary* for arriving at a certain result there is no question of relaxing or correcting the assumptions in the way we de-idealize cases like frictionless planes and so on; the abstractions are what make the model work (p. 110).

Morrison sharply contrasts two kinds of models that are typical of scientific endeavor: idealization is the more intuitive one, and it occurs when a “model idealizes or leaves out a particular property but allows for the addition of correction factors that bring the model system closer (in representational terms) to the physical system being modelled or described” (p. 111). The ladder of idealization is very easy to individuate and to climb up and down, and I feel like suggesting that many models we rely on in non-scientific practice partake of this nature: easy representational schemes for instance, maps, etc. allow us to perform inferences on idealized systems, and being able to perform these inferences is automatically associated with the ability to opportunely transfer the results to the target system.

Conversely, another kind of modeling – abstraction<sup>21</sup> – plays a conceptually pivotal role in scientific endeavor: it is the process of model, as I already said, by which the target phenomenon is essentially explained and constructed as such. This

<sup>20</sup> It is the same activity of *making sense of signs* that I described in section 2.2.

<sup>21</sup> Morrison connotes abstraction with a different meaning than [Woods and Rosales, 2010b] do. Their distinction between abstraction and idealization is comprised by Morrison’s definition of idealization.

is especially the case for models that allow a massive mathematization of the target system.<sup>22</sup>

[...] abstraction (typically mathematical in nature) *introduces* a specific type of representation that is not amenable to correction and is necessary for explanation/prediction of the target system. What is crucial about abstraction, characterized in this way, is that it highlights the fact that the process is not simply one of adding back and taking away as characterized in the literature; instead it shows how certain kinds of mathematical representations are essential for explaining/predicting concrete phenomena (p. 112).<sup>23</sup>

Morrison's example of this kind of mechanism is Maxwell's theory of electromagnetism, the development of which required a new model that supported the mathematization and the application of already known concepts, and this could not be worked out of idealization processes: "the foundation for electromagnetism emerged from the molecular vortex model and was in fact determined by it. But the important issue here is not that Maxwell was capable of deriving a set of field equations from a false model, but rather what it was about the model that underscored the applicability of the equations." Another example, the equations explaining the occurrence of phase transition in thermodynamics, had to be developed on similar models representing physically unrealizable situations, which are "required to explain a physically realizable one" (p. 130).

What these examples aim at showing is that to stress and investigate the fictional nature of scientific models is to look at the finger pointing at the moon and not at the moon itself, and – as usual – to ignore the fundamentally dynamic nature of scientific practice.<sup>24</sup> the construction of abstract models (which can even reverberate in the concrete exploitation of mediating artifacts, as showed by Faraday's experiments in the discovery of the first metallic colloid [Tweney, 2006]) plays a fundamental role in determining the target system itself. Resemblance, as stressed by [Magnani, 2012] in this book, cannot be used as a value guiding the development of the model (and the failure to comply with it a reason to judge the model as fictional): this is the case because resemblance is instituted aprioristically inasmuch as the phenomenon is individuated by the model that describes it, in a mutual engagement fitting with the idea of epistemic warfare. Then it can also turn out that the model does not resemble the target at all, by this does not necessarily cause the failure of the abstracting model, inasmuch as it receives some valuable

<sup>22</sup> The constitution of the target through the model is not a matter of developing strategically useful fictional accounts: "Introducing a mathematical abstraction that is necessary for obtaining certain results involves a different type of activity than constructing a model you know to be false in order to see whether certain analogies or similarities can be established" [Morrison, 2009], p. 111].

<sup>23</sup> Even though mathematization is the most straightforward example of creation of meaning subsequent abstracting modeling, other kinds of attributions of meaning exist: consider for instance Darwin's models of natural selection, which supported a significant new amount of meaning and individuated new features of the target even if not resorting to a massive use of advanced mathematics.

<sup>24</sup> The problem of the relationship between static and dynamic conceptions of science, relating to fictionalism, is analyzed by [Magnani, 2012] in this book.

feedback from the target system (e.g. accurate prevision, consistence with other models etc.)<sup>25</sup>

As I suggested at the beginning of this subsection, also on the basis of previous arguments, this *gnoseologically poetic* dimension of scientific modeling is indeed shared by emerging modeling which, sometimes, acquires a creative force in the assessment of external reality, especially when they set the ground for low-leveled abstractions such as mathematical ones. Of course, as I will show in the next and final subsection, this is not to say that emerging modeling and scientific models are exactly the same, but – as I have already contended in the introduction – it might be suggested that a fundamental difference does not originate from the nature of models themselves but from the attitude by which models are conceived and used in scientific practice.

### 3.3 *From Emerging Models to Scientific Models*

My aim in this last part of the paper will to suggest that part of the impetus of the Scientific Revolution resided in the new attention that was given to modeling, conferring them a new function (hence a new status) that allowed models to better relate to (and individuate) the laws of nature that science would aim at discovering. The concept of *epistemic warfare* will be pivotal to understand this claim.

If we frame the question in the argument so far, it appears that what is at stake is clearly not the invention of models, but of scientific models as we know it. My claim is therefore twofold: Galileo was aware that his models were indeed a prepared, and preparing, description (first claim) that supported the application of an advanced inferential systems and language, such as mathematics (second claim). It is interesting to read, with this project in mind, one of Galileo's most famous quotations:

In Sarsi I seem to discern the firm belief that in philosophizing one must support oneself upon the opinion of some celebrated author, as if our minds ought to remain completely sterile and barren unless wedded to the reasoning of some other person. Possibly he thinks that philosophy is a book of fiction by some writer, like the Iliad or Orlando Furioso, productions in which the least important thing is whether what is written there is true. Well, Sarsi, that is not how matters stand. Philosophy is written in this grand book, the universe, which stands continually open to our [p. 238] gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth [Galilei, 1957], p. 237–238].

Before we carry on, it is important to bear in mind that Galileo was referring to natural philosophy (which would be known as science) as a whole, and not specifically to scientific models. Therefore, out of honesty, the quotation should not be used *sic et simpliciter* as an authority weapon against those who advocate the fictional nature of scientific models: we would be committing nothing but a false implicature

---

<sup>25</sup> Consider what already stressed in footnote [17](#)



and a straw man fallacy, since to contend that scientific models are fiction does not coincide with affirming that the whole scientific endeavor has a fictional nature.<sup>26</sup> Nevertheless, as I am about to contend, the non-fictional character of scientific models is necessarily implied by Galileo's conception of natural philosophy.

Until the scientific revolution, natural philosophy would mostly perpetuate received models, which had a chiefly descriptive function: this was the essence of the Aristotelian "science" (the theory of natural places, for instance, would rely on descriptive, idealizing models that would provide a simplified vision of external reality). But the new conception of natural philosophy (i.e. science) could not be satisfied with models that were after all just arguable descriptions, favoring intrinsic qualitative and not quantitative analysis. The newborn science, in order to become intrinsically different than a "book of fiction" – where truthfulness is not a fundamental character for the appreciation of the work itself – had to rely on the construction of models that could grasp and produce an actual relationship between the model and the target system (external reality), even at the price of constructing this resemblance: scientists had to make the first move in the epistemic warfare against nature, and could not wait for nature to "amaze" them and direct their research, as prescribed by Aristoteles.

If we leave to science a chiefly descriptive function, then we witness the rise of two connected problems: on the one hand, models are prone to be nothing but descriptive accounts to be matched with a metaphysically rich external reality (it would be a mistake to call it phenomenon in this case), and hence always prone to be found fictional inasmuch as there is no clear criterion to define their truthfulness – the result is that truthfulness can be accounted on the bases of authority, especially when it is coupled with apparently self-evident truths, as in the case of Aristoteles affirming that lighter bodies fall slower. On the other hand, descriptive models as building blocks of a passive, descriptive science do indeed contribute to making science appear as a *book of fiction*, not because of its relationship with external reality, but because theories and models would be decided by likes and authority, as it was indeed the case when Galileo's heliocentric model was refuted inasmuch as it would endanger the Church's authority in the interpretation of the Scriptures.

Conversely, Galileo's conception of natural philosophy is that of an active quest, a true epistemic warfare: external reality begins to acquire its full dignity as a cluster of "phenomena", appearances where the self-evidence is not necessarily self-truthfulness, and a new conception of model is necessary. We need a kind of model that is conceptually poetic, that is to say, able to produce new phenomena by understanding and isolating them through a conceptually creative attribution of new meaning (connected to the discovery of new features).

In sum, the newly conceived model can be used to explain reality going beyond the simple received appearance, and this process is not the production an even

---

<sup>26</sup> [Giere, 2009] could be occasionally seen as slipping towards this fallacy, but it must be understood – as stated by Giere himself – as the will to preserve the dignity of science which has now to face extra-epistemic adversaries such as post-modern nihilism, aggressive creationism etc.: these actors are all too happy to commit the inverse fallacy and argue that if scientific models are fictional then science is fictional as a whole.

grander fiction (if I cannot trust my senses, how can I trust a model, when it is even *further* distant from reality?), because the fruitfulness of the abstracting quality of the model will be accounted for by the coherence of modeled (part of) reality itself, and the cluster of phenomena it entails. As I tried to show along this paper, emerging modeling in organisms – and especially in human beings – does indeed possess an abstracting nature which goes beyond the “simple” idealization, but in the XVI century this attitude was for the first time brought to full awareness and used as such, a clear example of which is Kepler’s discovery that the orbits of planets consisted in elliptical shapes [Gorman, 1998]: it was with an intentional act of poetic conceptualization that Kepler modeled the data he disposed so that they would fit in a novel geometrical pattern, and only (conceptually) subsequently this model pattern could support a successful mathematization, in the form of the ellipse. In this sense, therefore, I claim that for the first time Galileo acknowledged that models had to be used as prepared and preparing descriptions [Cartwright, 1983], also to the aim of actively delineating the phenomenon out of external reality.

Within this conception we can fully understand the experiment as the counterpart of the model as two (theoretically) distinct stages of epistemic warfare. In the first stage, i.e. modeling, the scientists carry out their “attack” on nature; in the second stage, the experiment, scientific endeavor stages a “passive” disposition where, in the typically controlled environment, the natural phenomena is allowed to strike back and test the value of the model (that is, behaving as assumed by the model). Without the experiment, the poetic abstracting nature of the model would condemn science (and other modeling activities) to be nothing but a solipsistic delirium.

In this sense, the experiment acquires its fullest meaning: it is not a game, something to impress other people and to show one’s skills, but a selective manipulation of a controlled environment that is artificially structured so to approximate the prepared description of reality embodied in the model. The experiment becomes therefore mutually bond with the model that had inspired it: the model affects the experiment and the experiment influences the model, together they manage to affect even the perception of reality.

Galileo’s new attitude towards the model is emphasized by the development of models as bald as the *thought experiment*, which could be seen as the *bootstrap*<sup>27</sup> phase of epistemic warfare: the scientist models the phenomenon (and thus isolates it), and then sets off the next stage by enacting nature’s response always within the model itself. [Gendler, 1998] shows how – even if Feyerabend would beg to differ – Galileo’s abstract modeling of a target system into a thought experiment was not the

<sup>27</sup> My use of the concept of *bootstrap* is similar to Nersessian’s as she contends that: “[...] the cognitive-historical method is the kind of bootstrapping procedure commonly used in science. The customary range of historical records, notebooks, diaries, correspondence, drafts, publications, and artifacts, such as instruments and physical models, serves as the source of empirical data on the scientific practices. The practices thought to be significant to the objectives of the analysis (in our case, creating concepts) are examined with respect to their cognitive bases. [...] The cognitive science research pertinent to analyzing the scientific practices comprises a wide range of investigations into how humans reason, represent, solve problems, and learn” [Nersessian, 2010], p. 6–7].

mere reproduction of a sophisticated but non-experimental argument: had this been the case, we would be back in another kind of metaphysical / theological modeling and have a merely doxastic reach [Faust, 2008]. Conversely, Galileo's model of the fall of two strapped bodies is structured to persuade "Aristotelians" as well, in a way that lets them persuade themselves.

The thought experiment that Galileo presents leads the Aristotelian to a reconfiguration of his conceptual commitments of a kind that lets him see familiar phenomena in a novel way. What the Galilean does is provide the Aristotelian with conceptual space for a new notion of the kind of thing natural speed might be: an independently ascertainable constant rather than a function of something more primitive (that is, rather than as a function of weight). It is in this way, by allowing the Aristotelian to make sense of a previously incomprehensible concept, that the thought experiment has led him to a belief that is properly taken as new [Gendler, 1998, p. 112].

The mental experiment can be rightly seen as bootstrapping the relationship between the model itself and the phenomenon it constructs and the reverberation of the experiment: it will of course require its enactment to surge to the status of a regular, physical experiment, but it plays nevertheless a fundamental role in the *épistémologie spontanée* embedded in Galileo's endeavor, that coincides with the spirit of epistemic warfare.<sup>28</sup>

This new conception of the model is so powerful that it has to bend reality (which is effectively reduced to a phenomenon depending on the model) which ultimately recovers human beings' emerging way of making sense of their experience. [Feyerabend, 1993] provides an interesting hermeneutic of Galileo's lexicons, and captures how the mathematical model inverted the order of dignity between model and observation to the point of reducing appearances to mere fallacy against more counterintuitive truths.

The senses alone, without the help of reason, cannot give us a true account of nature. What is needed for arriving at such a true account are "the... senses, *accompanied by reasoning*". Moreover, in the arguments dealing with the motion of the earth, it is this reasoning, it is the connotation of the observation terms and not the message of the senses or the appearance that causes trouble. "It is, therefore, better to put aside the appearance, on which we all agree, and to use the power of reason either to confirm its reality or to reveal its fallacy" [Feyerabend, 1993, p. 57].

Feyerabend also stresses how, despite their eventual success, the scientist's initial claims are far from being evenly proved, but Galileo "uses *propaganda*. He uses psychological tricks in addition to whatever intellectual reasons he has to offer. These tricks are very successful: they lead him to victory. [...] They obscure the fact that the experience on which Galileo wants to base the Copernican view is nothing but the result of his own fertile imagination, that it has been invented" [p. 65].

I should be able here to vindicate my second claim: it must be remembered, in fact, that as I showed in subsection 3.2 abstracting models are what support

<sup>28</sup> The relevance of this peculiar mental experiment to the ongoing debate about models is analyzed further and in greater detail by [Magnani, 2012] in this book.

the subsequent mathematization of the phenomenon (within and by the model) [Morrison, 2009]. Therefore, Galileo's ambitious claim that the book of the universe "[...] cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it," indeed points out to this *book* being constructed as a conceptual blend out of epistemic warfare.

It is necessarily a forced post-Kantian interpretation, but – on a careful reading – Galileo's inauguration of modern science does not seem to be echoing a Pythagorean conception of nature and its investigation, by which the laws emerge naturally and the scientist must just receive them. The newborn method was a conscious rationalization of how emerging modeling faculties could be turned into weapons to be used in an epistemic warfare between scientists and nature, in which the two opponents would *necessarily* taint each other but possibly in a virtuous way. The scientific intuition about the book of nature written in mathematical alphabet is that this is not a metaphysical given, but something acquired and projected by scientific endeavor.

On a provocative tone, it could almost be suggested that Galileo's was the first major successful attempt to lead philosophy to what could be called an eco-cognitive dimension [Magnani, 2009], that is appreciating the non-dissolvable theoretical connection between cognizant agents and their ecology, that is to say the environment on which their cognitive faculties operate: if scientific models are indeed a self-aware and rationalized successor of emerging natural models, then natural philosophy was *naturalized* indeed (since it would recognize the continuous bond between the philosopher and the natural framework she investigates), and could finally give birth to modern science in the same conception we have now of it.

## 4 Conclusion

In brief, this paper aimed at contributing to the ongoing epistemological debate on the nature of model by proposing an excursus from emerging modeling to scientific modeling that would highlight the similarities between spontaneous forms of modeling and scientific modeling: this analysis would also allow the rise of those traits that are instead typical of scientific models.

The analysis of basic forms of modeling tried to show how even mindless processing of external reality does not provide passive descriptions but is rather a poetic aggression which constitutes external reality as the organism perceives it. In my argument I foreshadowed several times how this poetic character is indeed common to both emerging and scientific modeling.

I took the liberty to resort massively to a concept introduced by [Magnani, 2012] in this book, "epistemic warfare": this concept is very useful to understand the difference between static and dynamic conceptions of science.<sup>29</sup> Paradoxically, to

<sup>29</sup> [Lockhart, 2008] contended that a static description of science is not recommendable even for didactical purpose as it completely spoils the nature, and thus the appeal, of scientific endeavor (Lockhart's contention specifically focuses on mathematics).

focus on the extent to which science can be seen as a warfare produces a double effect: on the one hand, it shows how the qualitative demarcation between the use of models in science and in the accomplishment of other cognitive tasks is often fuzzy, inasmuch as many prerogatives of scientific models (for instance their being constitutive prepared descriptions that support further inferential activity) are in fact widespread in model-based reasoning and can be said to be shared by basic model-driven activities such as perception. On the other hand, it shows how science is indeed characterized by a peculiar and conscious attention towards models as inaugurated by Galileo and the founding fathers of modern science. Such awareness of the power of the model was immediately paired with the other building block of science, that is the experiment (physical or mental), which served as a counterweight to the poetic virtue of the model as seen within the epistemic warfare: the experiment, coupled with the model, contributed to correctly locate nature's response and provisionally sanction the correctness of the model, so that the Baconian struggle is indeed a struggle between peers.

## References

- [Albertazzi *et al.*, 2011] Albertazzi, L., van Tonder, G.J., Vishwanath, D. (eds.): Perception Beyond Inference: The Information Content of Visual Processes. The MIT Press, Cambridge (2011)
- [Bertolotti and Magnani, 2010] Bertolotti, T., Magnani, L.: The Role of Agency Detection in the Invention of Supernatural Beings An Abductive Approach. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) Model-Based Reasoning in Science and Technology. SCI, vol. 314, pp. 239–262. Springer, Heidelberg (2010)
- [Bird and Emery, 2009] Bird, C.D., Emery, N.J.: Rooks use stones to raise the water level to reach a floating worm. *Current Biology* 19, 1410–1414 (2009)
- [Calvo and Keijzer, 2009] Calvo, P., Keijzer, F.: Cognition in plants. In: Baluška, F. (ed.) Plant-Environment Interactions: From Sensory Plant Biology to Active Plant Behavior, pp. 247–266. Springer, Heidelberg (2009)
- [Cartwright, 1983] Cartwright, N.: How the Laws of Physics Lie. Oxford University Press, Oxford (1983)
- [Cartwright, 2009] Cartwright, N.: If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis* 70, 45–58 (2009)
- [Cartwright, 2010] Cartwright, N.: Models: Parables v fables. In: Frigg, R., Hunter, M. (eds.) Beyond Mimesis and Convention. Boston Studies in the Philosophy of Science, vol. 262, pp. 19–31. Springer, Netherlands (2010)
- [Chakravartty, 2010] Chakravartty, A.: Informational versus functional theories of scientific representation. *Synthese* 172, 197–213 (2010)
- [Chandrasekharan, 2009] Chandrasekharan, S.: Building to discover: a common coding model. *Cognitive Science* 33, 1059–1086 (2009)
- [Clark, 1978] Clark, K.L.: Negation as failure. In: Gallaire, H., Minker, J. (eds.) Logic and Data Bases, pp. 94–114. Plenum, New York (1978)
- [Contessa, 2007] Contessa, G.: Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science* 74, 48–68 (2007)
- [Contessa, 2010] Contessa, G.: Scientific models and fictional objects. *Synthese* 172, 215–229 (2010)

- [Davis and Holmes, 2005] Davis, G., Holmes, A.: What is enumerated by subitization mechanisms? *Perception & Psychophysics* 67(7), 1229–1241 (2005)
- [De Cruz, 2006] De Cruz, H.: Towards a darwinian approach to mathematics. *Foundations of Science* 11, 157–196 (2006)
- [Dehaene *et al.*, 1999] Dehaene, S., Spelke, E., Pined, P., Stanescu, R., Tsivkin, S.: Sources of mathematical thinking: behavioral and brain imaging evidence. *Science* 284(5416), 970–974 (1999)
- [Dehaene, 1997] Dehaene, S.: *The Number Sense*. Oxford University Press, Oxford (1997)
- [Faust, 2008] Faust, J.: Can religious arguments persuade? *International Journal for Philosophy of Religion* 63, 71–86 (2008)
- [Feyerabend, 1993] Feyerabend, P.: *Against Method* (1975), 3rd edn. Verso, London-New York (1993)
- [Frigg, 2010a] Frigg, R.: Fiction and scientific representation. In: Frigg, R., Hunter, M.C. (eds.) *Beyond Mimesis and Nominalism: Representation in Art and Science*, pp. 97–138. Springer, Heidelberg (2010)
- [Frigg, 2010b] Frigg, R.: Fiction in science. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 247–287. Philosophia Verlag, Munich (2010)
- [Frigg, 2010c] Frigg, R.: Models and fiction. *Synthese* 172, 251–268 (2010)
- [Galilei, 1957] Galilei, G.: *The Assayer* (1623). In: Drake, S. (ed. & trans.) *Discoveries and Opinions of Galileo*, pp. 231–280. Doubleday, New York (1957)
- [Gendler, 1998] Gendler, T.S.: Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science* 49(9), 397–424 (1998)
- [Gibson, 1979] Gibson, J.J.: *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston (1979)
- [Giere, 1988] Giere, R.N.: *Explaining Science: a Cognitive Approach*. University of Chicago Press, Chicago (1988)
- [Giere, 2007] Giere, R.: An agent-based conception of models and scientific representation. *Synthese* 172, 269–281 (2007)
- [Giere, 2009] Giere, R.: Why scientific models should not be regarded as works of fiction. In: Suárez, M. (ed.) *Fictions in Science. Philosophical Essays on Modeling and Idealization*, pp. 248–258. Routledge, London (2009)
- [Glendinning, 2004] Glendinning, P.: The mathematics of motion camouflage. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271(1538), 477–481 (2004)
- [Godfrey-Smith, 2009] Godfrey-Smith, P.: Models and fictions in science. *Philosophical Studies* 143, 101–116 (2009)
- [Gorman, 1998] Gorman, M.E.: *Transforming Nature. Ethics, Invention and Discovery*. Kluwer, Dordrecht (1998)
- [Heyes, 1993] Heyes, C.M.: Imitation, culture and cognition. *Animal Behavior* 46, 999–1010 (1993)
- [Hodgson, 2009] Hodgson, D.: Evolution of the visual cortex and the emergence of symmetry in the acheulean techno-complex. *Evolution* 8, 93–97 (2009)
- [Holland, 1995] Holland, J.H.: *Hidden Order*. Addison-Wesley, Reading (1995)
- [Holland, 1997] Holland, J.H.: *Emergence: From Chaos to Order*. Oxford University Press, Oxford (1997)
- [Jacob and Jeannerod, 2003] Jacob, P., Jeannerod, M.: *Ways of Seeing: The Scope and Limits of Visual Cognition*. Oxford University Press, Oxford (2003)
- [Kirsh, 1995] Kirsh, D.: Complementary strategies: Why we use our hands when we think. In: Moore, J.D., Lehman, J.F. (eds.) *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pp. 212–217. Lawrence Erlbaum, Mahwah (1995)

- [Kuorikoski and Lehtinen, 2009] Kuorikoski, J., Lehtinen, A.: Incredible worlds, credible results. *Erkenntnis* 70, 119–131 (2009)
- [Laurent, 2003] Laurent, É.: Mental representations as simulated affordances: not intrinsic, not so much functional, but intentionally-driven. *Intellectica* 1-2(36-37), 385–387 (2003)
- [Lockhart, 2008] Lockhart, P.: Lockhart's lament (March 2008), [http://www.maa.org/devlin/devlin\\_03\\_08.html](http://www.maa.org/devlin/devlin_03_08.html)
- [Magnani, 2001] Magnani, L.: *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic/Plenum Publishers, New York (2001)
- [Magnani, 2004a] Magnani, L.: Conjectures and manipulations. Computational modeling and the extra-theoretical dimension of scientific discovery. *Minds and Machines* 14, 507–537 (2004)
- [Magnani, 2004b] Magnani, L.: Model-based and manipulative abduction in science. *Foundations of Science* 9, 219–247 (2004)
- [Magnani, 2006] Magnani, L.: Mimetic minds. Meaning formation through epistemic mediators and external representations. In: Loula, A., Gudwin, R., Queiroz, J. (eds.) *Artificial Cognition Systems*, pp. 327–357. Idea Group Publishers, Hershey (2006)
- [Magnani, 2007a] Magnani, L.: Animal abduction. From mindless organisms to artifactual mediators. In: Magnani, L., Li, P. (eds.) *Model-Based Reasoning in Science, Technology, and Medicine*, pp. 3–37. Springer, Berlin (2007)
- [Magnani, 2007b] Magnani, L.: *Morality in a Technological World. Knowledge as Duty*. Cambridge University Press, Cambridge (2007)
- [Magnani, 2009] Magnani, L.: *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Heidelberg (2009)
- [Magnani, 2012] Magnani, L.: Scientific Models are Not Fictions: Model-Based Science as Epistemic Warfare. In: Magnani, L., Li, L. (eds.) *Philosophy and Cognitive Science*. *SAPERE*, vol. 2, pp. 1–38. Springer, Heidelberg (2012)
- [Mäki, 2009] Mäki, U.: MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis* 70, 29–43 (2009)
- [Millikan, 2004] Millikan, R.G.: On reading signs: Some differences between us and the others. In: Kimbrough Oller, D., Griebel, U. (eds.) *Evolution of Communication Systems: A Comparative Approach*, pp. 15–29. The MIT Press, Cambridge (2004)
- [Mizrahi, 2009] Mizrahi, M.: Idealizations and scientific understanding. *Philosophical Studies* (2009), <http://www.springerlink.com/content/e33h421502t20118/>, doi: 10.1007/s11098-011-9716-3
- [Morrison, 2009] Morrison, M.: Fictions, representations, and reality. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 110–135. Routledge, London (2009)
- [Nersessian, 2010] Nersessian, N.J.: *Creating Scientific Concepts*. MIT Press, Boston (2010)
- [Pescic, 1990] Pescic, P.: Wrestling with proteus: Francis bacon and the “torture” of nature. *Isis* 90(1), 81–94 (1990)
- [Portides, 2007] Portides, D.P.: The relation between idealization and approximation in scientific model construction. *Science & Education* 16, 699–724 (2007)
- [Raftopoulos, 2001a] Raftopoulos, A.: Is perception informationally encapsulated? The issue of theory-ladenness of perception. *Cognitive Science* 25, 423–451 (2001)
- [Raftopoulos, 2001b] Raftopoulos, A.: Reentrant pathways and the theory-ladenness of perception. *Philosophy of Science* 68, S187–S189 (2001); *Proceedings of PSA 2000 Biennial Meeting*

- [Rivas and Burghardt, 2002] Rivas, J., Burghardt, G.M.: Crotalomorphism: a metaphor for understanding anthropomorphism by omission. In: Bekoff, M., Allen, C., Burghardt, M. (eds.) *The Cognitive Animal. Empirical and Theoretical Perspectives on Animal Cognition*, pp. 9–18. The MIT Press, Cambridge (2002)
- [Sage, 2004] Sage, J.: Truth-reliability and the evolution of human cognitive faculties. *Philosophical Studies* 117, 95–106 (2004)
- [Skelhorn *et al.*, 2010] Skelhorn, J., Rowland, H.M., Ruxton, G.D.: The evolution and ecology of masquerade. *Biological Journal of the Linnean Society* 99(1), 1–8 (2010)
- [Srinivasan and Davey, 1995] Srinivasan, M.V., Davey, M.: Strategies for active camouflage of motion. *Proceedings: Biological Sciences* 259(1354), 19–25 (1995)
- [Stevens and Merilaita, 2009] Stevens, M., Merilaita, S.: Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 423–427 (2009)
- [Suárez, 2009] Suárez, M.: Scientific fictions as rules of inference. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 158–178. Routledge, London (2009)
- [Suárez, 2010] Suárez, M.: Fictions, inference, and realism. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 225–245. Philosophia Verlag, Munich (2010)
- [Sugden, 2009] Sugden, R.: Credible worlds, capacities and mechanisms. *Erkenntnis* 70, 3–27 (2009)
- [Thomson-Jones, 2010] Thomson-Jones, M.: Missing systems and the face value practice. *Synthese* 172, 283–299 (2010)
- [Toon, 2010] Toon, A.: The ontology of theoretical modelling: models as make believe. *Synthese* 172, 301–315 (2010)
- [Tweney, 2006] Tweney, R.D.: Discovering discovery: How faraday found the first metallic colloid. *Perspectives on Science* 14(1), 97–121 (2006)
- [Weir and Kacelnik, 2006] Weir, A.A.S., Kacelnik, A.: A new caledonian crow (*corvus moneduloides*) creatively re-designs tools by bending or unbending aluminium strips. *Animal Cognition* 9, 317–334 (2006)
- [Weisberg, 2007] Weisberg, M.: Three kinds of idealization. *Journal of Philosophy* 104(12), 639–659 (2007)
- [Woods and Rosales, 2010a] Woods, J., Rosales, A.: Unifying the fictional. In: Woods, J. (ed.) *Fictions and Models: New Essays*, pp. 345–388. Philosophia Verlag, Munich (2010)
- [Woods and Rosales, 2010b] Woods, J., Rosales, A.: Virtuous Distortion: Abstraction and Idealization in Model-Based Science. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 3–30. Springer, Heidelberg (2010)
- [Woods, 2004] Woods, J.: *The Death of Argument*. Kluwer Academic Publishers, Dordrecht (2004)
- [Woods, 2010] Woods, J. (ed.): *Fictions and Models: New Essays*. Philosophia Verlag, Munich (2010)



# The Greenhouse Metaphor and the Greenhouse Effect: A Case Study of a Flawed Analogous Model

Xiang Chen

**Abstract.** Metaphors are double-edge swords. By connecting an abstract and unknown phenomenon to a tangible and familiar one, a metaphor also creates a new reality. For example, we frequently use a metaphor to describe global warming – the atmosphere works like a greenhouse and CO<sub>2</sub> traps heat as panes of glass in a greenhouse do. However, this greenhouse metaphor leads to an ontological assumption that conceptualizes heat as a material-like object, a series of ideas that ignore the roles of the ocean in the process of thermal transfer within the climate system, and an underestimation of the time delay effect in climate change. By producing an illusion that the climate system will respond instantly at the moment when CO<sub>2</sub> level is reduced, the greenhouse metaphor is ultimately responsible for the wait-and-see approach to climate change.

## 1 The Roles of Metaphors

By connecting abstract and unfamiliar subjects with tangible and familiar ones, metaphors play an important role in knowledge acquisition. However, metaphors typically highlight merely similar features between two subjects while ignoring many other dissimilar features. Rather than simply providing us with a way of conceptualizing abstract and unfamiliar subjects, metaphors actually create a new representation (Lakoff & Johnson 1980).

For example, global warming is the consequence of a very complicated process of energy exchange. After receiving energy from the sun, the Earth radiates the energy back into the space. Meanwhile, the atmosphere captures a part of the energy emitted by the earth's surface, which keeps the Earth warm. Since to understand this process of energy exchange requires many abstract notions that go beyond lay persons' perceptual experiences, a metaphor is widely adopted in science education and communication. People are told that the atmosphere works

---

Xiang Chen

Department of Philosophy, California Lutheran University, Thousand Oaks,  
CA 91360, USA

e-mail: chenxi@clunet.edu

like a greenhouse and gases such as CO<sub>2</sub> trap heat as panes of glass in a greenhouse do. This greenhouse metaphor helps us to understand an unfamiliar domain in terms of a familiar one. A greenhouse is a familiar and tangible object. Since panes of glass in a greenhouse and the atmosphere in the climate system have the same function of heat trapping, thermal transfer inside a greenhouse is in some aspects similar to energy exchange within the climate system. In this way, we can understand global warming through constructing a model for energy exchange in the climate system by mapping on it the process of heat movement inside a greenhouse (Vosniadou 1989). However, by highlighting certain features of the climate system while suppressing others, the greenhouse metaphor generates misconceptions inconsistent with scientific understandings of energy exchange in the climate system.

In the following sections, I first identify a few of these misconceptions. The greenhouse metaphor leads people to develop a framework to conceptualize heat as a material-like object. It leads people to develop beliefs that ignore the roles of the ocean in the process of energy exchange within the climate system. It also leads people to adopt a perspective that underestimates the effect of time delay in climate change. These misconceptions altogether produce an illusion that global temperature would be stabilized immediately at the moment when CO<sub>2</sub> level is reduced, and that we still have time to wait before taking actions. When people believe that the climate system would respond instantly to our adjustments, they may overestimate the time available for them to mitigate the problems of global warming. If people believe that the warming trend would be reversed immediately at the moment when we reduce the consumption of fossil fuels, it becomes reasonable to be cautious given the uncertainties in the research and it becomes logical to adopt the wait-and-see approach.

## 2 The Greenhouse Metaphor

The Earth receives energy from the sun. Roughly about a half of the energy from the sun is absorbed by the surface of the Earth. To maintain a balance, the Earth radiates the same amount of energy back into the space. Much of the outgoing energy emitted by the Earth is absorbed by the atmosphere (specifically, by certain constituents of the atmosphere such as CO<sub>2</sub> and water vapor), and then reradiated back to the Earth. In this way, the atmosphere keeps the Earth warm. It is estimated that the mean temperature of the Earth would be about 18 to 19 °C lower without the atmosphere.

The heat-keeping function of the atmosphere was discovered by the French physicist Joseph Fourier and the Sweden scientists Svante Arrhenius in the 19th century. Both Fourier and Arrhenius used metaphors to help people to understand the heat-keeping function of the atmosphere. Fourier compared the Earth covered by its atmosphere to a box covered with a pane of glass, and later Arrhenius used a “hot-house”, a metaphor that associates with lay people’s daily experience, to explain the functions of the atmosphere (Fleming 1998). Since then, the metaphor of a hot-house, or a greenhouse, had become an intrinsic part of the theory accounting for the process of thermal transfer among the sun, the Earth and

the atmosphere, and the heat-keeping function of the atmosphere had been called the greenhouse effect.

Today the greenhouse metaphor is widely used in science education and science communication to illustrate the process of thermal transfer during climate change. We can find a typical example of the metaphor in use from the EPA website that aims at children. The website begins with an introduction of the greenhouse effect, calling it “the rise in temperature that the Earth experiences because certain gases in the atmosphere trap energy from the sun.” Then, the website offers an image of a greenhouse, and explains that the glass panels of a greenhouse let in light but keep heat from escape. Finally, it concludes that “greenhouse gases in the atmosphere behave much like the glass panes in a greenhouse. ... Some of the energy passes back into space, but much of it remains trapped in the atmosphere by the greenhouse gases, causing our world to heat up” (USEPA 2009).

The greenhouse metaphor is influential, shaping students’ and lay persons’ understandings of the process of thermal transfer in the Earth. Studies investigating people’s conceptions of the greenhouse effect reported that only a small percentage of subjects have a scientific understanding of the thermal transfer process. Among those who attempted to account for the warming trend, they frequently mentioned the greenhouse effect, but they gave a literal interpretation of “greenhouse”, as a place where heat is trapped to raise plants (Shepardson, et al. 2009).

Although both the atmosphere and a greenhouse are similar because of their heat-keeping function, they actually operate according to different principles. What happens inside a greenhouse is a process of thermal convection – glass panes of a greenhouse form a “blanket” that completely cuts off the upward heat flow. However, the atmosphere heats up the Earth through a process of thermal radiation. Unlike a greenhouse, the atmosphere forms merely a partial “blanket” – only a few components of the atmosphere (CO<sub>2</sub>, water vapor and several other trace gases) can absorb the outgoing thermal energy. Unlike a greenhouse, the atmosphere absorbs the outgoing energy of radiation selectively – only the long-wave radiation is stopped. Also unlike a greenhouse, the atmosphere reemits energy to all directions – only a part of the absorbed energy is sent back to the Earth. Thus, CO<sub>2</sub> and other “greenhouse” gases warm the Earth in a manner quite different from the way through which a greenhouse warms its interior. Strictly speaking, it is inappropriate to call the heat-keeping effect of the atmosphere a “greenhouse effect” because, as we will see in the following sections, the greenhouse metaphor generates confusions about climate change.

### **3 The Nature of Heat**

The greenhouse metaphor compares the atmosphere with panes of glass in a greenhouse. Since a hole in the glass panes of a greenhouse would change its interior temperature, a hole in the atmosphere could also affect global temperature. When a “hole” in the ozone layer was discovered in the 1980s, it became almost logical for many people to relate the ozone hole to the warming phenomenon,

mistakenly believing that the ozone hole can somehow be a cause of global warming.

Interviews of children aged 13 to 14 years reveal several explanatory models underneath the belief that regards ozone holes as one of the causes of global warming. A dominant model is that holes in the ozone layer allow more sun rays or heat rays to enter the atmosphere. Since these heat rays cannot find the holes to escape, the Earth heats up and we have global warming. Another model holds that ozone holes allow more ultraviolet rays to enter. Since ultraviolet rays are “hotter” than other kinds of heat ray, the Earth heats up. There is another model believing that ozone holes allow air to escape. Since the higher the altitude the colder the air, more cold air escapes and the Earth becomes warmer (Boyes & Stanisstreet 1997).

Because the greenhouse metaphor is dominant, the belief that regards ozone holes as one of the causes of global warming is widespread, especially in the early discussions of climate change. In the 1990s, 95% of the general public believed that stratospheric ozone depletion is the cause of global warming (Bostrom, et al. 1994). The confusion about ozone holes is also persistent. After having been informed that the greenhouse effect is the cause of global warming for more than a decade, many people continue to associate ozone holes with the warming trend (Khalid 2003).

The persistency of the confusion between ozone holes and the causes of global warming indicates that it is probably associated with some fundamental assumptions about the nature of heat. For those who consider ozone holes as the cause of warming, they somehow assume that a hole is needed for heat to penetrate. This is a reasonable assumption if heat is a material-like object. From life experience we know that objects always occupy space and their occupancy is both unique and exclusive. Because objects occupy and compete with space, two objects cannot occupy the same place at the same time, and one particular place cannot be occupied by more than one object at a given time. Thus, when heat is understood as a material-like object, it is logical to assume that heat always occupies space exclusively, and that an open space or a hole in the atmosphere is needed for heat to pass through. This assumption about the nature of heat generates a robust belief that ozone holes are responsible for the warming trend. When the sun is considered as the source of thermal energy, it is logical that a “hole” in the atmosphere would allow more sun rays to come in and make the Earth warming.

The idea that heat is a material-like physical object is also intrinsic to the greenhouse metaphor. A greenhouse keeps its interior warm by preventing heat from leaving through thermal convection, the transfer of heat by the actual movement of warmed air. When air is warmed, it expands and rises, carrying thermal energy with it. What happens inside a greenhouse is a mechanical process, in which heat is carried and dispersed by observable movements of air. When we analyze thermal convection within a greenhouse, we focus on the movement of air, a three-dimension object that inherits many properties of materials. Thus, when we discuss thermal transfer within a greenhouse, it is not only appropriate but also necessary to adopt a framework that treats the subject of analysis as a

material object. In this way, the greenhouse metaphor implies an ontological assumption about the nature of heat, that the subject of thermal transfer is a material-like object.

#### **4 The Role of the Ocean**

The notion of heat as object results in several misconceptions of the process of thermal transfer within the climate system. When heat is understood as an object, convection, that is, the process of observable motions of warmed matter, becomes the only type of thermal transfer. Without the notion of thermal radiation, it becomes impossible to comprehend the process in which outgoing thermal energy emitted by the Earth is first absorbed by the atmosphere and then reradiated back to the Earth. Within the framework conditioned by the greenhouse metaphor, the thermal function of the atmosphere is regarded merely as a convective insulator that stops the flow of heat current. Also, when heat is understood as an object, it becomes impossible to comprehend the phenomena of thermal radiation, which are processes that involve transfer of energy through electromagnetic waves. Especially, it becomes impossible to comprehend that all bodies with a certain temperature emit electromagnetic radiation and that bodies with different temperatures emit energy with different frequencies. Within the material framework of heat, the role of the ground is regarded merely as a thermal reflector.

These misconceptions together prevent us from understanding the roles of the ocean in the process of thermal transfer within the climate system. When heat is understood as an object, heat is associated with hotness in proportion to temperature – the hotter an object, the more heat it can emit. In other words, the level of hotness of an object is understood as an intensive property, a variable that does not depend on the size or the amount of material in the object. Thus, the material account of heat has a very limited picture of the objects involved in the process of thermal transfer within the climate system. It only considers objects that are relatively hot, such as the sun and probably also the atmosphere, as the sources of thermal energy. Objects that are relatively cold, such as the ocean, never appear in the picture. But heat or thermal energy is in fact an extensive variable – how much thermal energy an object contains also depends on the size of the object or the amount of material in the object. The ocean is a very important link in the process of thermal transfer because of its size. Because the ocean covers 71% of the Earth's surface and because heat can be stored to a depth over 1,000 meters, the ocean can absorb a large amount of heat – the ocean has about 1,000 times the heat capacity of the atmosphere. However, when heat is understood as an object, it becomes very difficult, if not impossible, to appreciate the important roles of the ocean in climate change.

The popularity of the material notion of heat explains why the general American public has little knowledge of the roles of the ocean in climate change. The majority of the general public have a perception that climate change is a terrestrial phenomenon. They see the ocean simply as a source of moisture that plays no roles in heat transport. Particularly, they fail to understand, due to the

huge heat capacity of the ocean, climate change has to be observed not in annual but in decadal time scales (The Ocean Project 2009). The ignorance of the important roles that the ocean plays in the process of thermal transfer within the climate system is one of the cognitive factors responsible for the wait-and-see approach on climate change.

## 5 The Illusion to Instant Responses

When the roles of the ocean are ignored, thermal exchange in the climate system is in many ways similar to water accumulation in a bathtub, a model used frequently to represent the dynamics of a simple system. A bathtub is a simple system of stock and flow, in which the level of a single stock (water) is determined directly by two flows – the amount of water coming in from the faucet (the inflow) and the amount of water going out through the drain (the outflow). It seems that the level of thermal energy stored in the atmosphere (the stock) is also determined by two flows – the energy that comes from the sun (the inflow) and the energy that emits into the space (the outflow). Thus, it seems that the climate system can also be treated as a simple system with the atmosphere as the only storage of energy. In such a simple system, the change of the stock is determined by the net flow – the difference between the inflow and the outflow. In a bathtub, the level of water rises when the inflow is larger than the outflow. Similarly, it seems that the level of thermal energy in the climate system follows the same principle, that is, the level of thermal energy in the atmosphere and the global temperature increase when the energy inflow is larger than the energy outflow.

From life experience, we know that we can stabilize the water level of a bathtub at the moment when we shut off the water supply. The response of the system to our action is instant without delay. If we treat the climate system as a simple system similar to a bathtub, it is reasonable to expect that we can stabilize the amount of thermal energy in the atmosphere at the moment when we balance the energy outflow with the inflow. In other words, it is reasonable to expect that we can stop the warming trend at the moment when we bring the rate of energy outflow back to normal by reducing the concentration of CO<sub>2</sub>.

However, the climate is not a simple system of stock and flow. Warming is the direct consequence of the accumulation of thermal energy in the atmosphere. However, thermal energy stored in the atmosphere is not the only stock responsible for the warming result. In addition to an accumulation of thermal energy, there is also an accumulation of CO<sub>2</sub>, which determines the amount of thermal energy to be emitted into the outer space (the outflow). These two processes of accumulation occur not only in the atmosphere, but also in the ocean and biosphere. The ocean is a huge storage for both heat and CO<sub>2</sub>. The ocean can absorb both heat and CO<sub>2</sub> from the atmosphere and transfer them to its interior through various physical, chemical and biochemical processes. Heat and CO<sub>2</sub> stored in the interior of the ocean will later return to the atmosphere.

Because of the ocean's huge capacities in absorbing heat and CO<sub>2</sub>, the exchanges of heat and CO<sub>2</sub> between the atmosphere and the ocean are slow. It takes many years for both heat and CO<sub>2</sub> to reach the deep ocean, and many more

years for them to return to the atmosphere. The exchanges of heat and CO<sub>2</sub> between the atmosphere and the ocean must be observed with a multi-decadal scale (Solomon, et al. 2007). Thus, even when the movements of heat and CO<sub>2</sub> from the atmosphere to the deep ocean declines, the movements from the deep ocean to the atmosphere would continue at a high rate because the current climate has generated a large amount of heat and CO<sub>2</sub> stored by the deep ocean. It will take decades of time for the movements of heat and CO<sub>2</sub> from the deep ocean to slow down.

Because of the huge absorptive capacities of the ocean, the stock of heat in the climate system does not respond instantly to the difference between the energy from the sun and the energy to the space, nor does the stock of CO<sub>2</sub> respond instantly to the difference between CO<sub>2</sub> emitted by human activities and absorbed by nature. There are time lags between adjustments in a net flow and a change of a stock. The climate system behaves in a way similar to how a heavy object behaves in a mechanical system. Because of its mass, a heavy object has a tendency to resist changes in velocity and to maintain its current state of motion. In physics, we use the notion of inertia to describe and understand such a resistance. As a metaphor, we can also say that the climate as a complex system also has inertia, a resistance to change in the current conditions of the processes of accumulation.

However, when people treat a subject as a simple system, it becomes difficult for them to appreciate the existence of inertia or the effect of time delay. After they initiate an adjustment or a control action, they expect the system to respond instantly. Consequently, their adjustments or control actions lead to overshooting behaviors of the system – adjustments that aim to correct problems in the system do not prevent the problems from getting much worse before they start to have effects (Moxnes 1998).

Such an illusion to instant responses also exists in people's perception of the dynamics of the climate system. In a series of studies, Sterman and Booth Sweeney found that many people, including highly educated (graduate students at MIT), treat the climate as a simple system and expect it to respond instantly to adjustments of the energy inflow and outflow (Sterman & Booth Sweeney 2007). In their studies, Sterman and Booth Sweeney asked the subjects to estimate the changes of global temperature in several hypothetical scenarios, including one in which human CO<sub>2</sub> emissions suddenly fall to zero and one in which the concentration of CO<sub>2</sub> in the atmosphere is reduced to a level slightly lower than the current one. Because the climate is a complex system with inertia, it would not respond to these adjustments instantly. Cutting CO<sub>2</sub> emissions down to zero would not stabilize CO<sub>2</sub> concentration immediately, and reducing CO<sub>2</sub> concentration would not lower global temperature right away. According to one estimate, after CO<sub>2</sub> emissions drops to zero, global temperature would continue to rise for about three decades before it goes down (Fiddaman 1997). However, a majority of the subjects in Sterman and Booth Sweeney's studies failed to understand the delay effect. More than a half of them mistakenly believed that global temperature would drop or stabilize immediately after anthropogenic CO<sub>2</sub> emissions drop to zero, and a majority of them mistakenly thought that global temperature would

stabilize in a few years after the concentration of CO<sub>2</sub> in the atmosphere is reduced.

The illusion to instant responses has important policy implications. To those who treat the climate as a first-order linear system as a bathtub, there are reasons to believe that we still have time to deal with the environmental crisis. As we can stabilize the water level in a bathtub by waiting until the last moment when the water reaches a particular threshold, we can also stabilize global temperature by waiting until the temperature reaches an unbearable level. Without the concept of inertia in the climate change, it is reasonable to be cautious, and to adopt an approach to wait until we have more evidence. In this way, the illusion to instant responses is directly responsible to an overestimation of our perceived adaptation ability and a failure in recognizing the urgency in responding to the climate crisis. To those who believe that reducing CO<sub>2</sub> concentration can immediately stabilize global temperature, there are indeed no needs at this moment to make any immediate and costly responses (Chen 2011).

## 6 The Need for a Conceptual Change

The greenhouse metaphor plays an indispensable role in teaching and informing the greenhouse effect to the public. Heat and thermal transfer are abstract phenomena that cannot be comprehended on the basis of direct observations. To understand the scientific concepts of heat and thermal transfer we need theories of thermal dynamics. It is very difficult for people who have not studied physics to understand these phenomena correctly. To overcome the difficulty, we need a metaphor that connects the abstract phenomena to people's daily experience. Such a metaphor should be tangible, offering concrete objects for people to simulate the phenomena under study. The greenhouse metaphor serves this purpose, offering such concrete objects as glass panes and material-like heat for people to simulate the greenhouse effect.

But the greenhouse metaphor is incomplete in the sense that it simulates only a part of the mechanisms behind climate change. Since a greenhouse works by preventing heat from leaving through convection, the greenhouse metaphor does not depict thermal transfer by radiation, the mechanism that actually keeps the Earth warm. Because it fails to capture radiation, the greenhouse metaphor is not a good model to simulate the mechanisms responsible for global warming and climate change.

Even worse, the greenhouse metaphor is flawed by depicting heat as a material-like object. Ontologically speaking, heat is not an object but a process, a succession of changes that take place over time. Specifically, heat is the transfer of kinetic energy between objects with different temperatures. When the greenhouse metaphor depicts heat as an object, such an assumption is reinforced by one of our inner cognitive biases. Studies show that people have a tendency to prefer objects over processes, or to view processes as objects. This preference originates from a unique sequence of our own cognitive development – we cannot distinguish processes from objects until the age of seven, but we have already developed a core system of object knowledge as early as 4 months of age (Keil 1979;



Baillargeon, et al. 1985). Children's understanding of objects that developed at the beginning of their lives constitutes an important part of our core systems of knowledge, and adults never completely give up this core system of object knowledge (Carey 2009). The core system of object knowledge is well entrenched, and people prefer to rely on this core system of knowledge whenever it is possible. This preference for objects is called the "object bias" (Chen 2007). Because of the object bias, people tend to ignore the ontological differences between the thermal transfer in a greenhouse and that in the climate, and treat heat as a mechanical object. In this way, the incomplete metaphor of a greenhouse becomes a flawed metaphor, generating a series of robust and persistent misconceptions inconsistent with climatic sciences. Specifically, the greenhouse metaphor reinforces a material notion of heat, which results in ignorance of the ocean in the process of energy exchange, and an underestimate of the effect of time delay in climate change. These misconceptions altogether produce an illusion that we still have time to wait before taking actions.

Thus, the greenhouse metaphor is a double-edge sword. It offers a helpful model to bridge abstract physical knowledge with daily experiences, but it also generates misconceptions responsible for the wait-and-see approach toward global warming. It is unrealistic not to employ the greenhouse metaphor in the discussion of climate change, but we need to find a way to communicate with the public correctly and effectively.

To correct the misconceptions caused by the greenhouse metaphor, we need a general discussion of the related ontological assumptions. We need to communicate with the public the fundamental differences between "object" and "process" as two distinct ontological categories. Specifically, we need to alter the flawed ontological assumption about the nature of heat. Since the material notion of heat exemplifies a general cognitive bias, the preference to treat various ontological entities as objects, we need a conceptual change, a fundamental transformation from an object-only perspective to a perspective that properly treats objects and other kinds of entities, particularly processes, as distinct kinds. This is a conceptual change between lateral categories, that is, categories on different branches of the hierarchical tree of our conceptual system (Chi 2008). More specifically, this is a transformation from object concepts to process concepts with many characters different from transformations between concepts belonging to the same ontological kind (Chen 2010).

Conceptual changes that cross different ontological categories are not only necessary to correct the misconceptions associated with the greenhouse metaphor, but also frequently needed to the applications of metaphors in general. In practice, most metaphors that we adopt in both natural and social sciences are systems of objects operating according to mechanical principles, simply because objects are tangible. When we use mechanical models to simulate entities that belong to different ontological categories, misconceptions are bound to emerge. To correct these misconceptions, we need conceptual changes to shake off the object bias and subsequently to adopt a proper ontological perspective.

## References

- Baillargeon, R., Spelke, E., Wasserman, S.: Object permanence in 5-month-old infants. *Cognition* 20, 191–208 (1985)
- Boyes, E., Stanisstreet, M.: Children's models of understanding of the two major global environmental issues (Ozone layer and greenhouse effect). *Research in Science & Technological Education* 15, 19–28 (1997)
- Carey, S.: The origin of concepts. Oxford University Press, Oxford (2009)
- Chen, X.: The object bias and the study of scientific revolutions: Lessons from developmental psychology. *Philosophical Psychology* 20, 479–503 (2007)
- Chen, X.: A different kind of revolutionary change: Transformation from object to process concepts. *Studies in History and Philosophy of Science* 41, 182–191 (2010)
- Chen, X.: Why do people misunderstand climate change? Heuristics, mental models and ontological assumptions. *Climatic Change* 108, 31–46 (2011)
- Chi, M.: Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In: Vosniadou, S. (ed.) *International Handbook of Research on Conceptual Change*, pp. 61–82. Routledge, New York (2008)
- Fiddaman, T.: Feedback complexity in integrated climate-economy models. Ph.D Thesis, MIT Sloan School of Management, Cambridge MA 02142 (1997)
- Fleming, J.: Historical perspectives on climate change. Oxford University Press, New York (1998)
- Keil, F.: Semantic and conceptual development: An ontological perspective. Harvard University Press, Cambridge (1979)
- Khalid, T.: Pre-service high school teachers' perceptions of three environmental phenomena. *Environmental Education Research* 9, 35–50 (2003)
- Lakoff, G., Johnson, M.: *Metaphors we live by*. University of Chicago Press, Chicago (1980)
- Moxnes, E.: Overexploitation of renewable resources: The role of misperceptions. *Journal of Economic Behavior & Organization* 37, 107–127 (1998)
- Shepardson, D., Niyogi, D., Choi, S., Charusombat, U.: Seventh grade students' conceptions of global warming and climate change. *Environmental Education Research* 15, 549–570 (2009)
- Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., et al.: *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge (2007)
- Sterman, J., Booth Sweeney, L.: Understanding public complacency about climate change: Adults' mental models of climate change violate conservation of matter. *Climate Change* 80, 213–238 (2007)
- The Ocean Project, America, the ocean, and climate change: New research insights for conservation, awareness, and action (2009), <http://theoceanproject.org/resources.php>
- US Environmental Protection Agency (2009), *Climate change: Kids site*, <http://epa.gov/climatechange/kids/greenhouse.html>
- Vosniadou, S.: Analogical reasoning as a mechanism in knowledge acquisition: a developmental perspective. In: Vosniadou, S., Ortony, A. (eds.) *Similarity and Analogical Reasoning*, pp. 413–437. Cambridge University Press, New York (1989)

# A Study of Model and Representation Based on a Duhemian Thesis

Chuang Liu

**Abstract.** This paper contains four lines of inquiries based on Duhem's thesis about the contrast between the abstract French mind and the concrete British mind. The first line brings out the real contrast between two types of methods and their results: the A(bstract) method or models and the C(oncrete) method or models. The second line gives a critical discussion of the Callender-Cohen deflationism on scientific representation. The third discusses Russell's structuralism in connection to the theme. And the fourth critically discusses the relationship between models and fiction in connection to the distinction between the A-models and the C-models. A conclusion maps out the author's view on the nature of the C-models and why they, and only they, can be viewed as fully fictional.

## 1 The Thesis of Duhem

“To a Frenchman and a German,” Pierre Duhem, a profoundly philosophically minded French physicist, wrote before the turn of the twentieth century in a now famous text whose English translation bears the title of *The Aim and Structure of Physical Theory*,<sup>1</sup> “a physical theory is essentially a logical system. Perfectly rigorous deductions unite the hypotheses at the base of a theory to the consequences which are derivable from it and are to be compared with experimental laws. ... Thus the French and German founders of mathematical physics, the Laplaces, the Fouriers, the Cauchys, the Amperes, the Gauses, the Franz Neumanns, have constructed with extreme caution the bridge intended to connect the point of departure of the theory, the definition of the magnitudes it is to deal with, and the justification for the hypotheses which will bear its deductions to the road on which its algebraic development will proceed.” (Duhem 1954, p. 78)

---

Chuang Liu

Center for Phil of Science and Technology, Shanxi Univ., China

Dept of Philosophy, Univ of Florida, USA

Center for Science, Technology, and Society Studies, Tsinghua Univ., China

e-mail: logics@ufl.edu

<sup>1</sup> The translation was made from the 2<sup>nd</sup> edition of the book which was published in 1914 as *La Théorie Physique: Son Objet, Sa Structure*, by Marcel Rivière & Cie., Paris.

In stark contrast, Duhem continued, “[u]nderstanding a physical phenomenon is, . . . , for the physicists in the English school, the same thing as designing a model imitating the phenomenon; whence the nature of material things will be understood by imagining a mechanism whose performance will represent and simulate the properties of bodies. The English school is completely committed to the purely mechanical explanations of physical phenomena.” (p. 72)

And furthermore this profound difference between the English and the Continental ways of theory construction was seen as a species of a more general difference between two types of minds – the *deep and narrow* versus the *ample and shallow*. “The opposition between the French mind, strong enough to be unafraid of abstraction and generalization but too narrow to imagine anything complex before it is classified to a perfect order, and the ample but weak mind of the English will come to us constantly while we compare the written monuments raised by these two peoples.” (p. 64)

For examples, Duhem compared Shakespeare’s plays with the classical French theatre noting the chaos and heterogeneity of plots and dialogues in the former versus the formal rigor and elegance of the same in the latter. In physics, while Maxwell’s genius in constructing such models as the phantom ‘displacement current’ in dielectric materials, which is an imitation of Ampere’s real current in conductors, was acknowledged and even admired, William Thomson’s, or Lord Kelvin’s, many mechanical ‘inventions’, such as one finds in his magisterial treatise *Lectures on Molecular Dynamics*, were regarded as distasteful monstrosities; and most repugnant of all were the many mechanical models for the aether that particularly excited and exercised the English mind, which to Duhem was a clear demonstration of degenerate method if not degenerate mind.

What is remarkable upon a closer reading of this portion (chapter IV of part I) of Duhem’s book is that there are no clear-cut examples of the products of a Continental mind. It was acknowledged that the best of the Continental thinkers also build models, and yet theirs are either not nearly numerous or not as grotesquely in mechanical details as the English ones. This raised the question of whether Duhem was against building models in theory construction at all or he was just against building mechanical models, which follows from a general distaste for the philosophy of mechanisticism that was thought to have mostly been brought to high fashion by the British minds in the late nineteenth century. Perhaps implicit in Duhem’s observations is a division of ‘good models’ versus ‘bad models’.

## 2 Methods Rather Than Minds

It would be easy and entirely justified to dispute and even reject Duhem’s characterization of the English or the French mind; even Duhem himself qualified such a crude dichotomy by giving examples of some of the best French minds engaging in model building, e.g. Descartes was no less imaginative than Maxwell in constructing his vortices in cosmology and Napoleon’s mind was terribly ample but shallow. And yet, Duhem may well be re-interpreted as trying to distinguish two types of minds, regardless of whether they could be neatly instantiated in two

separate nations or regions. The fact that no actual male body exemplifies pure masculinity, or no actual female body pure femininity, does not refute the distinction between the two opposite dispositions. But it still seems problematic to think that a specific way of doing physics strictly corresponds to a type of mind. It does not seem plausible that such a claim can have any credibility in contemporary psychology. However, when Duhem uses the term ‘mind’ to denote an English or a French mind, he is perhaps best understood *not* as talking about it in psychological terms or in terms of what today’s philosophers of mind would be permitted to use. Instead he may be best interpreted as contrasting two alternative (scientific) methods that are frequently adopted by minds of two different dispositions.

One method of constructing and evaluating a theory for a given phenomena in physics is aimed at representing the phenomena in highly abstract terms and discovering general principles or lawful propositions among the terms. Let’s henceforth call this method the A method, for being abstract or axiomatic. From the abstract propositions, rigorous logical and mathematical derivations may be given whose end-products may then be brought into comparison with experimental generalizations. The method does not so much condemn model-building practices as shuns them; and even if models are built and used in such a method, they are models of abstraction as to allow further distancing from concrete and actual systems or events, rather than models of ‘flesh and blood,’ i.e., models whose physical embodiment plays an indispensable role in *explanation* and *prediction*, two of the main aims of science.

The other method – let’s call it C method, for being concrete – encourages and values model-building activities as essential to scientific representation. It is not that this method uses any less mathematics than the A method does; the difference lies chiefly in the ways in which the phenomena are *represented*, *ontologically regarded*, and *explained*.

Here is how Duhem is regarding the difference of representation. For the A method, the representation is entirely abstract: measurable magnitudes are defined and symbolized and functional relations among such magnitudes or time-evolution equations of these, or sets of these, magnitudes are given. What these magnitudes ‘embody’ as physical systems are not important and only left to our faculty of imagination which according to Duhem has little to do with science. In contrast, the C method uses representations that are concrete systems of mostly observable parts, and how concrete it goes depends of course on what types of models are called for: some types such as hydraulic models of air flows are more so than others such as the model of a stock market. How should the phrase “concrete systems of mostly observable parts” be understood in general is, on the one hand, crucial to the distinction between the two methods and difficult to unpack with complete satisfaction, on the other. I shall return to this point in Section 4 when I discuss Russell and structuralism.

As I mentioned above, the chief difference between the two methods is how the phenomena is represented, ontologically regarded, and explained. So far, the C method doesn’t seem to have any advantage over the A method. Can we always regard the models with concrete details ontologically real, namely, telling us what

the represented systems really are? Such questions are notoriously difficult to answer, not only as general philosophical questions but also as historical ones. Did astronomers defending the Ptolemaic system believe ontologically in the reality of all the epicycles? Did Lord Kelvin believe ontologically in his convoluted ether models? Do physicists today believe ontologically in the 'ghosts' in the Higgs mechanism? A more poignant case in point: we know that the superstring theory of elementary particles is supported by a mathematical theory that as mathematics is mature and sophisticated so that the kind of logical rigor that Duhem demands as the quality of a Continental mind is in ample evidence. The problem is with the model of the strings, which is baffling if it is taken as a concrete physical system. But isn't there a perfectly sound Duhemian argument against questioning the superstring theory because one cannot conceive of a plausible model for the strings?

If anyone gives up superstring theory just because its concrete model appears baffling and does this consistently for every newly proposed theory, she is justifiably condemned by the Duhemians; but we like most scientists are wiser than that. To regard a concrete model as real does not commit us to the overly naïve stance that says the real thing is exactly like the model. A more reasonable view that is consistent with realism is to (1) believe that the model refers to something real that exists independently of how we think of it and (2) think that the model resembles *approximately* the real thing, where what degrees of resemblance counts as acceptably approximate depends on a number of pragmatic factors that include the demand of explanation/understanding and the expectation of the discipline and (3) believe that the degree of approximation can always be improved as we know more about superstrings.

Here a lesson from Locke may be noted. Locke thought that our ideas of secondary qualities of external bodies are caused by the secret powers of those bodies that are forever concealed from us. We may have genuine knowledge of the bodies but not their secret powers because by the basic principle of empiricism we can only know them through secondary qualities (together with ideas of primary qualities on the macroscopic scale). Such a view can no longer be held because we no longer have to depend on our impressions of color or warmth or texture in order to know about an object. In order to know, for instance, the color and warmth of a certain surface, we only need to measure the frequencies of the reflected electromagnetic waves and the frequencies of vibration of the molecules on the surface. Whether or not these frequencies are indeed the intrinsic properties of the waves and the surface may be a more profound philosophical question, but there is no denying that Locke is wrong to think that just because we cannot observe the 'secret powers' that produce the secondary qualities which do not belong to the bodies themselves, we are forever barred from knowing what those powers are. Similarly, just because what the superstrings represent cannot 'look like' spacetime strings does not entail that no improved models will tell us what they really are.

As for explanation, is it true then that a theory made by using the A method simply *cannot explain*?<sup>2</sup> It is obviously not true; an abstract theory may provide just as much causal explanations of a certain phenomenon as the concrete theory. It all depends on what kind of explanation is expected, and sometimes too much concrete details hinders rather than aids the effort of explanation (cf. Bokulich 2004). Including the size and shape, not to mention all the surface attachments, of the earth does not help at all if what we want explained is why it revolved around the sun in the observed orbit. However, there are also many cases of explanation in which an abstract theory is simply not adequate. Imagine trying to explain the weather phenomena by using an abstract theory of air mass, treating clouds and such merely by their geometric size and shapes and their mass, or trying to explain divorce in a society by using an abstract theory of econometrics (where humans are represented as perfectly rational agents).

From this analysis we can see that no real progress can be made in truly understanding these two methods unless we go deeper into some more troubled water in philosophy of science, such as the nature of modeling, of scientific representation (in general), and of structuralism, to which we now turn, beginning with the second variation in which the two methods are examined in the light of today's conception of models and modeling.

### 3 Models and Representation

Models are often regarded within or without philosophical literature as representational devices, and scientific theories, which may or may not include models, must be able to represent before they can be used for other purposes, such as to explain and predict. And how well theories can provide explanations and predictions must depend crucially on how well they represent the phenomena in question. Part of Duhem's complaint about models being superfluous derives from his belief that one does not need models to fulfill the aim of science, namely, prediction (if not explanation). Although he does not explicitly mention or argue for it in his book, Duhem no doubt believes that theories can represent without models.

A version of this attitude, albeit with more sophistication, was recently expressed in an article by Callender and Cohen (Callender & Cohen 2006), in which is given what I shall call a 'deflationary' concept of scientific representation. The upshot of the Callender-Cohen idea is that models don't have

---

<sup>2</sup> In Chapter 1 of the book (pp. 7-18), Duhem has argued against the appeal to explanation in physics. The matter is unfortunately complicated by the historical context in which this argument by Duhem is made. The target is 'metaphysical explanation' in Duhem's term, and the examples mostly involve using occult qualities for explanation. It is almost obvious that if today's notion of scientific explanation, such as the one broached by Hempel, is used, Duhem would not have any dispute with it. But again, Hempel's models of scientific explanation might be faulted as insufficient for such a notion precisely because it eschews its metaphysical implications.

to be ‘models’ in order to fulfill their representational role<sup>3</sup> (where the models we have been considering so far are taken to be artifacts, material or mental, that show a *resemblance* of some sort to the represented); anything, any device, can be a model, as long as it successfully represents the system it is intended for. Callender and Cohen in a sense resolved the Duhem problem concerning the use of models in science by re-conceiving what models may be as representational devices. As we noted earlier, even in the A method, a system has to be represented in some way before a theory can be conceived about it; and right there models are employed in the Callender-Cohen deflationary sense.

Callender and Cohen suggest in their paper that much confusion in the literature comes from trying to provide answers for the wrong questions: a case in point: people have been trying to figure out in what general and minimal sense a model could be said to resemble its target – whether it be similarity or isomorphism – while addressing the question of what it is that *constitute* the relationship between the two. This mistake, they argue, is caused by confusing the ‘constitution question’ about models or modeling in science with the ‘demarcation problem’ whose solution demands some sort of criterion for distinguishing those representational devices that can from those that cannot serve specific purposes (similarity is obvious important if the model is created to provide a visual representation of the target). And this problem should be further distinguished, according to Callender and Cohen, from what they termed the ‘explanatory/normative problem’ of scientific models and modeling, which looks for answers to questions such as ‘what makes a model the *correct model* for a given phenomenon,’ or the question ‘in virtue of what do models represent and how do we identify what constitutes a *correct representation*?’ as they quoted from a paper by Margaret Morrison (Morrison 2006, my italics).

These distinctions are long overdue, and one couldn’t help recalling a similar situation in the theory of truth. The inquiry into what truth is and how it can be found had been regarded one of the most profound and difficult undertakings in the history of philosophy until Tarski came along and offered his deflationary idea of what truth is, namely, the scheme of ‘*s* is true if and only if *s*. However, this answer to the constitutional question of truth does nothing to tell us which sentences in a given domain are true and which are not, nor does it show which true sentences are important to us and which are trivially true.

The basic idea of the Callender-Cohen deflationary theory is that scientific representations are *derivative* representational devices that are reducible to the *fundamental* devices, which are kinds of mental states that people (e.g. members of a scientific community) agree by convention to use as chosen tags for the target phenomena/systems. What these mental states are may be a deep metaphysical or

---

<sup>3</sup> Models in Duhem’s conception are obvious a small proper subset of the models that are currently regarded as such. Originally, by models Duhem really meant mechanical models; but even if we loosen this restriction and take them to be more than mechanical models, Duhem’s models are still much closer to the common sense conception of scientific models than any contemporary philosophers’ models. The context of discussion in this paper should make it clear which notions of models we are talking about in a particular instance.



scientific question of cognition that does not have to be answered, or might have controversial answers, before we know exactly what models are and how they represent. Specific Griceanism is where Callender and Cohen begin and it is basically a reductive account of how linguistic tokens function as representation of objects or states of affairs in reality. Words or sentences on paper or other surfaces which are intended as linguistic or graphic representational devices do their job by being related to mental states that reliably appear in a person's head when what they represent is intended to be reflected or communicated. Scientific representational devices, such as models, do their job in accordance with General Griceanism, which is a natural extension of Specific Griceanism to representational devices in general. The basic scheme of representation is of course the same, and it gives a unified account of how any derivative representational devices do their job in representing the world to us. To aid their arguments, Callender and Cohen mentioned such acts of representation as lanterns being raised in a certain way at a certain hour to represent the presence or absence of enemy troops, or more dramatically, salt shaker on your dinner being used to represent your favorite geographical region Madagascar (Callender & Cohen 2006, 13-14). The key and only condition of adequacy is that the right intentional states are invoked among the users of the devices.

Neither Specific nor General Griceanism, according to Callender and Cohen, can or needs to spell out what kind of fundamental devices we must have in order for us to have successful representation. One only needs a plausible argument for the reduction of the derivative devices to the fundamental ones and for ensuring that the nature of reduction does not in principle put any constraints on what types of entities may serve as models. The conclusion is that on this aspect of General Griceanism, little if anything more needs to be said beyond what is already said in Specific Griceanism about linguistic devices. As long as symbolic markings or objects are used to evoke corresponding mental states that by convention reliably produce understandings of the represented, anything can be a model.

In this respect, Teller (2001) can also be viewed as holding a similar deflationary theory for models and modeling. Regarding what scientific models should be, he says.

I take the stand that, in principle, anything can be a model, and that what makes a thing a model is the fact that it is regarded or used as a representation of something by the model users. Thus in saying what a model is the weight is shifted to the problem of understanding the nature of representation (Teller 2001, 397).

And so perhaps to a less straightforward sense does van Fraassen think of representation when he observe that if one is to have a theory of representation (which he doesn't) one must accept what he takes as the '*Hauptsatz*': "*There is no representation except in the sense that some things are used, made, or taken, to represent some things as thus or so.*" (van Fraassen 2008, 23). And for any object or state of affairs, what things so ever cannot be *used* to represent it if stipulated by a convention? So, anything can be used to represent and what counts as correct representation is a matter of convention (i.e. pragmatics).

And if this is right, if anything can be used to represent anything as long as the right mental states are invoked, such hotly debated notions as the ‘similarity’ between the model and the modeled or the ‘isomorphism,’ or some even fancier terms such as ‘partial isomorphism,’ between the two have nothing to do with the ‘constitution’ of scientific representation that we call ‘models;’ (see also Suárez 2003 on this point).

### *Two Separate Questions*

Be that as it may, I still see a fundamental problem being evaded here. When we ask how scientific models represent, we may be taken as asking one of the following two distinct questions. We may be asking

(1) What kind of devices is appropriate for us to use in representing the world around us?

Or

(2) How do we represent the world around us?

Question (1) is answered by the deflationary theory while question (2) is not. Or perhaps I should say that in order to answer question (2), we need to know at least some general constraints on the fundamental representations (which are supposed to be in our head). In other words, the question about scientific models or representation is not only (or really) a question about what external devices we can use to represent but also (or rather) a question about what can be the content of our mental states when we act to represent. It is at least about what general constraints need to be placed on such contents.<sup>4</sup>

To put this point slightly differently which may highlight the difference, we could say that the question about scientific representation (models and modeling) is about (a) how we can put in material or visual forms what fundamental representations we have in our mind and more importantly (b) how we can put in material or visual forms what an extended fundamental representations we have in our mind. The extended fundamental representations are those for things we cannot directly perceive, such as atoms or electromagnetic field. And I suggest that our scientific models for such things are material or symbolic replications of what we ‘see’ in our mind’s eye of what they really are, or at least what scientists would like to have us see. It is certainly not true in this sense that any devices, words, lanterns, or gestures, would do, as the deflationary theory claims.

From the deflationary theory we get: anything can be used to represent what we want to represent *as long as it evokes the right kind of mental states in our each other’s mind*, but what does that mean? What could be the content of such mental states? First of all, it must contain the belief that accomplishes the reduction as mentioned above. Whatever it is, and it could be of various kind, the

---

<sup>4</sup> A further question is of course a central question in philosophy of mind about what exactly happens when we perceive the world around us. Do we perceive it primarily pictorially? There must be judgments mixed in but then how do it work? For answers to these and many other related questions, see for instance, Siegel 2011, in which a Rich Content View of perception is defended. See also, Freeman 1991.

belief has to be of the effect (which one must recognize in one's mind) that the device refers to the object that it is agreed upon to pick out. When my community agrees to use three coconuts hanging above the front door to signal a medical emergency inside, a belief state must be evoked by that whose content must be something to the effect that the holder of that belief knows that there is a medical emergency inside that door when seeing three coconuts hanging above it. Secondly, it also contains something else, something in my mind that allows me to connect the device with, an *image* of that which the device is used to represent, perhaps? Can that be anything we want, just as what the deflationist would say a model can be?

When the problem we should be investigate is "how do we (humans) represent the external world around us?" rather than "what we can use as props or marks to represent something else in the world around us?" the deflationary theory is no longer adequate. The question about models and modeling may well be taken as concerning the former rather than the latter problem. Callender-Cohen and Teller are right to argue that derivative representations must be reducible to the primary ones in order to work, and the primary representation is done in our head, but they are wrong in thinking that realizing the reductional relation is all that is needed to solve the constitution problem of models and modeling in science. Scientific representation in the form of modeling is not aimed at coming up with symbols or objects that help us to bring out what is in our head or what *should be* in our head that ultimately represents; it is rather aimed at finding appropriate material or symbolic rendering of that primary representation in our head.<sup>5</sup>

Wittgenstein once proposed in his *Tractatus* (Wittgenstein 1961) a picture theory of meaning, and we find statements such as "A proposition is a picture of reality. A proposition is a model of reality as we imagine it (4.01)." "One name stands for one thing, another for another thing, and they are combined with one another. In this way the whole group – like a *tableau vivant* – presents a state of affairs (4.0311)." Now if it is up to Callender and Cohen or Teller, regarding propositions (if they accept the existence of propositions, which they are unlikely to do) as pictures of reality may be overreaching and unnecessary but to regard them as 'models of reality' is surely unproblematic. In general, as far as representing the world around us is concerned, the picture theory of meaning should be acceptable to the deflationists.

Now here is a simple example that ought to be entirely unproblematic for Wittgenstein and the deflationists, and yet it does not seem quite right intuitively. What makes a word in English, such as 'water', represents water must be the same as what makes an object (serving as a model as the word means in common-sense), such

---

<sup>5</sup> There is in all this a big and fundamental epistemic assumption, which some epistemologists may not accept; and that is that we represent the world around us primarily by images or impressions in our head. This assumption goes back at least to Descartes and is the basis for British empiricism as in Locke and Hume. It is possible not to accept this assumption and think that we directly perceive objects outside without any mental representation of them (such as in Reid's direct realism). It would be difficult to account for images in memory with such a position but it is not an impossible position to hold.

as a plastic array of water molecules denoted by 'W', represents it. In other words, 'water' is a model of water in the same sense W<sup>6</sup> is. And if Wittgenstein is right, 'water' is a picture just as W is one, and both are no less pictures of water than a photo of a glass of clear water. They may differ in their pragmatic roles in representing the thing, and yet as far as constitutional question is concerned, they are all pictures/models of water.

Wittgenstein's picture theory of meaning has been widely criticized and I doubt any philosophers today still believe in it. Looking at one particular criticism briefly may help us to see more clearly the difference between the two questions (1) and (2) mentioned a few pages back. In a critical discussion of 'The Picture Theory of Meaning,' E. Daitz (Daitz 1956) argued that Wittgenstein's theory couldn't be right because of some fundamental differences between words and sentences on one side and pictorial objects, such as painting, sculpture, and mechanical models, on the other, in terms of how they represent what they do.

Pictorial representations in general, which Daitz called 'iconic,' contain elements that represent corresponding elements in the represented; and in addition, although the elements do not have to individually bear any resemblance relations, the connection among the representing elements must bear certain perceptually identifiable resemblance to the relationship among represented elements; and it is this latter feature that distinguishes this type of representation from what Daitz called 'purely conventional' representations, to which linguistic ones belong. To use our example given above, the word 'water' is elemented by 'w', 'a', 't', 'e', 'r', which do not correspond to any element of the represented, namely, to any element of water, and moreover, the concatenation of these letters show no resembling relation to the chemical bound that connect those elements. On the contrary, the molecular model W of water is a typical example of the iconic representation, where each distinct part of the plastic model correspond, in however a rough and ready way, a water molecule, and how those plastic parts are put together is supposed to resemble the chemical bounding of the water molecules.

Therefore we can say, and there is no question that Daitz is right on this point, that we can represent an apple, an traffic accident, a scene in a play, or even an imaginary creature, such as Santa Clause, both with propositions and pictures, but propositional representations are fundamentally different from pictorial ones, even with respect to the most fundamental constitutional question of representation. Iconic representations and conventional ones are of two mutually exclusive (and perhaps jointly exhaustive) methods of representation. To argue for such an intuitively appealing claim one would have to figure how our mind ultimately represents the world around us. In other words, it is the question I raised earlier when criticizing Callender and Cohen, namely, we need to find out what constraints the primary mental states that do the representation in our mind. It is nothing less than part of the investigation of the mind-world relation.

---

<sup>6</sup> Notice the difference between 'water', which mentions the word in the sentence, and W, which uses the letter that stands for the molecular model of water. The question is how similar or different does the word represents water as W does.

Another of Callender-Cohen's thesis is that we can bracket the inquiry into the question of fundamental devices (which are appropriate mental states) when we investigate the various problems of scientific presentation. The constitution problem can be solved, as we saw above, through understanding the reductional relationship between derivative and fundamental devices (General Griceanism); and the demarcation problem concerning what types of devices are correct scientific models can be solved through pragmatic considerations such as matching the right type with the right needs in scientific explanation and/or prediction. Seeing in this manner, the question about whether similarity or isomorphism is in some sense indispensable in scientific representation should not be regarded as relevant to the constitution problem; and at most it might be relevant to the demarcation problem because it is really a matter of pragmatics.

This thesis, we should by now realize, is not quite right. If there is a sense, which by now we should have no doubt, in which the constitution problem is a problem about the fundamental devices, about how we represent in the basic sense, then it matters whether relations such as similarity or isomorphism are required. The reason for this point is very simple. Despite the widespread use of conventional devices by humans to represent whatever they want to represent such that it appears anything *can be used* to represent anything, our mind may ultimately only represent what we experience pictorially or non-pictorially or it represents some parts of reality pictorially but other parts non-pictorially. The very fact that it takes agreement within a community to use the conventional devices to represent reality means that the mind does not *naturally* use anything like those (e.g. impressions that resemble those) to represent reality.

I do not pretend that I can discuss here in high degree of care and clarity the question of how we represent as one of the most fundamental questions of the mind-world relation. But these observations should be safe to make. If our mind represents the external world around us mainly in iconic ways<sup>7</sup>: we see shapes and colors and so forth, and we hear sounds of different pitches, and we feel bodies of different textures, and such are indeed in the content of our mental states that constitute our experience<sup>8</sup>, then we have sufficient reasons to demand that our models of what we can experience directly (i.e. the observables) pictorially resemble what we see, hear, and/or feel. Conventional devices may have to be used when for one reason or another it is not a good idea to try to construct pictorial models. Bohr's model of hydrogen atoms comes to mind, and this is an issue I will return in the next two variations in which I discuss in turn structuralism and model's connection to fiction. However, if our mind did not represent the world chiefly in pictorial ways, then pictorial devices have to be regarded as conventional and reductions have to be carried out in order to connect such a device to the fundamental device that our mind used. Imagine that computers as they are made now (not some future supercomputers that might have

---

<sup>7</sup> I am here using 'iconic' or 'pictorial' to refer to qualitative images of all kinds in our experience; it is not only 'visual.' A more technical term for this might be Russell's 'percept' or 'perceptual.'

<sup>8</sup> There should be little doubt that our senses create in our mind pictorial images of what come through them; and it is also what we see assumed in the history of philosophy.

humanlike capacity of perception) have consciousness and are engaged in discussing this very same issue. Since they do not ‘experience’ the world around them by anything other than strings of binary characters, their fundamental device of representation might well be non-pictorial; and they represent the world primarily by propositions, which are not and cannot be pictures, as Diatz convincingly explained.

To summarize in a very blunt manner, Callender and Cohen are right in arguing that anything can be used to represent in science as long as it is what is needed and reducible to the right types of mental states; but they are wrong to assume that the question of what scientific models (or representational devices) should be has nothing to do with the fundamental question of how we represent. I suggest here that it does and because the way we represent is essentially iconic (or perceptual), questions of what types of models best fulfill the task is not a secondary question. It then explains why such relations as similarity or isomorphism are widely regarded as essential to scientific modeling; and such a widespread view, especially in the science community, is not a conceptual mistake.

#### 4 Models and Structure

“The French or German physicist conceives,” Duhem wrote in the chapter whose text I quoted in Section 1 to broach the theme, “in the space separating two conductors, abstract lines of force having no thickness or real existence; the English physicist materializes these lines and thickens them to the dimensions of a tube which he will fill with vulcanized rubber. In place of a family of lines of ideal forces, conceivable only by reason, he will have a bundle of elastic strings, visible and tangible, firmly glued at both ends to the surfaces of the two conductors, and, when stretched, trying both to contract and to expand...” (Duhem 1954, p. 70). Moreover, as he explained in more detail what a French or German physicist, “be he a Poisson or a Gauss” (p. 69), would do for the study of two conductors in space, Duhem described a method of idealization by abstraction: idealizing the two conductors into two point charges in empty space, and imagining the electric force acting along the 1-d line that connects the two points, etc. The rest, such as establishing the equation for the force and its effects on the movement of the point charges, and how such equations can be used to derive observable results, are, I assume, shared activities by both schools, or with any other school that is capable of producing a workable theory for electromagnetism.

In today’s conception of models and the model-building practice (Morgan & Morrison 1999, Hughes 1997), we would say that both Continental physicists and British ones are engaged in model building, the only difference is that they build different kinds of models. The French and the German, if Duhem is correct, which is a big ‘if’ that I shall not entertain, are accustomed to using the highly abstract models, while the British like to indulge themselves with concrete ones. To use the terms we defined in Section 2, the tension between the A method and the C method is a real tension but a tension within model-building methodology; and it is not just a matter of style, or so shall I argue in this and the next variation.

I want to suggest here that the tension is mostly created by differences of philosophical or semi-philosophical viewpoints, which appear not just among philosophers of science, and it is resolved among non-philosophers mostly by pragmatic considerations. With observable, macroscopic systems it seems that abstract models with varying degrees of abstraction are most appropriate. We hardly need any concrete models to represent such systems when they are there for us to 'see.' Such models are necessarily results of idealization in terms of abstracting away properties that are not pertinently related to the ones we study. With unobservable or microscopic systems, models of both types may be needed; and when the models are concrete ones, they are often the results of analogical reasoning. And in reasoning by analogy, structural similarities, such as isomorphism, plays a central role (see, Hesse 1966).

Let's take a look first of Russell's structuralism (cf. Russell 1927, 1959; Demopoulos & Friedman 1989; Hylton 1990; Demopoulos 2003a). Russell's structuralism is founded on a firm belief that there is a one-one onto mapping between the world of our experience and the real world; in other words, the two worlds are *isomorphic to each other*. Although cognitive agents like us only have access to their own experience, the sciences, especially in their theoretical parts, are according to Russell about reality, about facts in the realm of real events and objects – or just events, if objects are regarded as abstractions out of events – that exist independently of us. However, such scientific knowledge is always highly abstract, containing only propositions about how certain types of events or objects are related to certain other types in a lawful manner or how parts of an object are related to one another. And all these are characterized by values of variables that are only definable *structurally*, such as, for instance, the time, position, and momentum of a classical particle, where the first two express temporal and spatial relations, respectively, and the last a potential of motion or motion production that is also a relational magnitude. The reason for the sciences being so abstract and purely structural is, for Russell, the result of what we have to do to 'get beyond' our experience to 'reach' reality. We cannot reach reality as it is, as the thing-in-itself, to borrow a familiar Kantian phrase, directly or via perception because of the mediation of representations in our head; but because of the fact that there exists a structural identity between what exists around us and how it appears to us, which is guaranteed by the relation of isomorphism, we at least can know reality by its structures. From perception, we are acquainted with our surroundings first and foremost *qualitatively*, and yet our scientific knowledge can say nothing about what reality is like qualitatively, e.g. it is neither colored nor warm or cold nor loud or quiet nor textured; and therefore, in a sense, the real world is not knowable to us; or we have no right to say, for instance, that an apple is really a solid roundish object with juicy and flavorful flesh wrapped inside a smooth skin that is either green or red (or green or yellow) or a bit of both. However, we can know the structure – when it is understood as a catchall word for any kind of relational properties – of reality, or we know scientifically that an apple is a three dimensional object that is composed mostly of empty space, and in it molecules of various types that are related in a certain type of configurations – one type for its flesh and another for its skin, etc.. We don't know what molecules are

qualitatively, just as we don't know what apples are qualitatively, and yet we know how they are made up, i.e. in what configurations, by atoms and how they structurally make up bigger objects, such as apples. The structure vs. quality distinction is total for Russell, namely, we 'know' just about everything that appears to our senses twice over: once by its quality through our senses and then by its structural properties through science. If one were to invoke Locke and his theory of secondary qualities here, one might say on behalf of Russell that we do know about the 'secret powers' of things that produce impressions of secondary qualities, but we know them not as qualities that resemble the representations in our head; we know them as a bunch of numbers, such as frequencies of vibration.<sup>9</sup>

Russell argued for his structuralist position chiefly by noting the fact that causal chains that preserve perceptual qualities are often made up of radically different media. Just look at what happens in electronic communication chains, whether it is telephone or television. To go from what the producers of the records or movies are doing to the listeners or viewers of their products, vast distances are traversed by these products in environments which bear no resemblance to the ones in which they are produced and the ones in which they are shown, and yet incredible fidelity is maintained through the transmission. The best way to account for such success is to think that what is captured in the production and then transmitted and recovered for the listeners or viewers are the physical structures of the phenomena. And such structural properties are the only things we can know scientifically of reality and the only thing we need to know for the purpose of using them to explain how things work and events occur causally in reality.

Putting aside the question of whether or not Russell's structuralism holds, a question I shall discuss later in this section, I find it illuminating at this juncture to reflect on the above-mentioned tension between the two types of methods or models. If Russell is right, what science does and can present to us is only a 'world' of relations, which means it cannot tell us what the relata of such relations really are. If we do feel that we know from physics or other sciences what they are like, we do so by adding the perceptual to the structural, whether or not we are justified in so doing. Sometimes the added perceptual content in helping to construct a model is what we actually perceive, but more often we use what we imagine we might perceive to fill in the structural. All such addition would be regarded as erroneous according to Russellian structuralism.

---

<sup>9</sup> A brief word about the notion of structure and why structural similarities do not tell us about what things are like as themselves. A structure in Russell's sense refers to any relational property of an object (or event): how the parts of it are related to one another and how it is related to other relevant objects, etc. It turns out to be the same as a quantitative property, and therefore a structural similarity between two objects won't tell us what the objects are but that the relational properties of the two map in pairs 1-1 onto each other. It is in this way, and only in this way, Russell argued, that we know the cause of our perception, i.e. all the quantitative aspects of the object we perceive. When we see a red ball, we may know that it is produced by the arrangement and motion of parts of the ball to produce a certain frequency in the reflected light from its surface. Such a belief if true is only knowledge about relations in the ball and between parts of the ball's surface and the light beam that it reflects.



Should we then straightforwardly condemn all scientific models in which terms or images for sensible qualities are regarded as genuinely referring? Should we then regard Russell's structuralism as endorsing Duhem's condemnation of concrete models or the C method? That is not necessarily true and the reason is simple. Not all non-qualitative models have to be abstract in the sense that Duhem demands; or concrete models are fine as long as they are quantitative models. And it all depends on how one interprets the terms or images that are included in a model. Take Duhem's example of two charged objects interacting with each other and affecting each other's movement. The A model would have us imagine that the charged objects are point particles and the interacting force/field between them as lines of field/force, while the C model include such qualities as the sizes of the objects and the thickness of the lines of field/force. The latter are indeed sensible qualities for we do represent them as such in our head. However, they are also structural properties of the objects and field/force as well. The C model of this system does not have to be thought of as including sensible qualities, nor is the A model of this system the only model that is consistent with Russell's structuralism. We only abandon that position for certain if we also include color or hardness in our scientific representation of the charges or lines of force.

In the previous variation I argued for an anti-deflationary idea that says that scientific models are conceived as answers to questions of how we represent, and our representation of the world around us is primarily iconic rather than conventional. Now if we incorporate Russell's structuralism, I should say that the models can and should only be about the structure of the iconic representations in our head. No qualitative properties of models, physically built or otherwise, should be counted as relevant properties to their representational role.

Hence, even though Russell did claim that all our scientific descriptions are abstract, he meant by that word in a different and much more general way than Duhem did. The structures could be of all sorts when we construct models for reality, and the only limit is set by our perceptual experiences. We are all right so long as there is an isomorphic relation between the structure of our representation and the structure of that which is supposedly causing it. Salmon (1984) seemed to obviously have Russell in mind when he argued for his view on causal processes as structure carriers. And here is one of the obvious benefits if Russell is right. Structuralism means that a resemblance between a theoretical description and its target system can only be a mapping of two sets of relations; and the mapping is either a bijection or an injection or a surjection. Partial mappings (or partial functions where the mappings are 1-many or many-1) are also sometimes useful, but functions are by far the most common in structural representations of reality. Qualitative resemblance, a notion that besets the discussion of modeling and simulation, is therefore ruled out. For Russell, all qualitative stuff only exists in our mind, reality is only known through quantitative resemblances.

Van Fraassen (1980) also speaks of isomorphism between models and phenomena. For van Fraassen who espouses a version of empiricism, as oppose to scientific realism, that is called 'constructive empiricism,' what constitutes a

belief of a theory is not a belief of its being true but rather of its being empirically adequate or its capacity of 'saving the phenomena;' and the notion of empirical adequacy is cashed out, according to van Fraassen's semanticist standpoint on the nature of scientific theories, in terms of an isomorphism between the empirical sub-model of a theory's model and the phenomena, where the former are constructions in the realm of the observable. For instance, for a mechanical theory of motion, the theory that is couched in theoretical terms might be a set of differential equations, in which such items as instantaneous velocity or acceleration are not observables. But it is ultimately connected to trajectories of the moving objects in question, as solutions of those differential equations; and segments of trajectories are certainly observable. These trajectories are empirically adequate, and so is the whole theory by implication, if and only if they bear a one-one correspondence (i.e. an isomorphism) with what is observed in the labs in which such motions are studied.

But what does it exactly mean to have a 1-1 onto mapping between structural elements (which comprise an empirical sub-model of a theory) to phenomenal or qualitative elements (which presumably comprise our experience)? How can a sub-model of statistical mechanics about a bucket of water be isomorphic to our experience of the water's texture and coolness? If Russell is right, there couldn't possibly be any relationship between that model and our qualitative experience of the water in the bucket. What there is can only be a relationship between that model and the structural elements of our experience, i.e. the structure of the phenomenon, because the relationship is between the model and the cause of our experience and we can only know the cause structurally. What counts as the phenomenon makes a difference in this case. If the phenomenon is created by putting my hands into the water to feel its texture and coolness (or warmth), no isomorphism between such qualitative precepts (to borrow a term from Russell) and a sub-model of statistical mechanics is possible. We would have to invoke the 'structure' of the feelings from my hands in order to make sense of the isomorphism. But what is that structure? How do we find out about that structure? Here physics tells us roughly that our feeling of texture and coolness should be accounted for by the viscosity and the temperature of the water; and these can be measured by reliable instruments. When we read off the numbers from those appropriate measuring devices, we are ready to think about the isomorphism. But are the two sets of numbers, one for the viscosity and the other for the temperature, the proper stuff to element a phenomenon? One may say that such a problem does not exist for our previous example about a moving object since the observed trajectory is a phenomenon and can be regarded as isomorphic to the corresponding solution of the differential equations for the same object. But if we are talking about a precept, namely, a mental representation of a moving object, there is no 'observed trajectory' in our head. To get a trajectory, instruments have to be used and numbers recorded, so we are looking at the same situation as the water example as far as an isomorphism (or the lack of it) between two structures is concerned.

These considerations also show how sketchy Russell's original structuralism is. To say the cause of our perceptions must be structurally isomorphic with the

structure of the precepts (its effect) would not make good sense until we spell out how the structure of our precepts is known, which to say the least are the result of a complex combination of causal factors from outside which may possess extremely heterogeneous structures.

## 5 Models and Fiction

As devices of scientific representation, whether their uses are methodologically justified or not, can models be viewed as fiction or fictional objects? There is a recent flourish of discussion, most of which comprises various defenses of the positive thesis, but the defenses are mostly in the spirit of trying to enrich our understanding of the nature of models and modeling, and so the arguments are mostly along the line of “there seems to be some similarity between models and fiction, so let’s explore the relationship between the two and see what it tells us about models.”

In this section I shall explore this relationship along the line of the philosophers who recently broached the subject (see Frigg 2010, 2011a, 2011b and Suarez 2009a, 2009b), but I shall be more critical. My aim will be to bring objections to the conceptions of this relationship so that eventually we see in what exact sense we are entitled to say that models in science are like characters or events in a fiction.

Objections to regarding scientific models as fictional objects should be easy to find. Science and fiction do not mix in our common sense conception of the two. Scientific theories are for the real and the actual while fiction depicts by default the opposite. When something in science is called ‘a fiction,’ it usually means that it is unworthy of science. And science fiction as a genre of fiction is not, and cannot be regarded as, a discipline in science. Imagine the surprise if one finds in a bookstore the section of ‘science fiction’ on the same wall for science books right next to such sections as ‘physics,’ ‘biology,’ and ‘psychology.’

Frigg (2010, 2011a, 2011b) rightly separates two senses of ‘fictional.’ One means simply something untrue or nonexistent, while the other means the result of imagination. When people call something in science ‘fictional,’ the context in which the term is used demands that the first but not the second sense is meant. When people called cold fusion, for instance, a piece of fiction, they were not thinking how imaginative that ‘discovery’ is. This is certainly an important distinction, and yet perhaps it can and should be made a bit more accurate. One sense of being said to be fictional is not so much of being false as of being non-referring, of not having anything in reality to which the imagined item, an object or an event, correspond. The other sense is not so much about something being *a result of* imagination as about it *being in* imagination. When we call something fictional in this sense, we intend to stress that it is a thing in the mental realm or it is not in space and time or it is not and cannot be individuated as a particular.<sup>10</sup> *Prima facie*, if models in science are understood as fictional in the first of these two senses, the understanding is flawed, while it is presumably harmless to regard

---

<sup>10</sup> Here I use the metaphysical theory of individuation by Peter Strawson (1959).

them as fictional in the second sense. Nobody would think of models in science, though they sometimes have particulars as their embodiment, as themselves particulars. This, I think, clearly explains why the material realizations of a model are usually ‘immaterial.’

Another point concerning models and fiction needs to be clarified right away. It is one thing to say some models in science are fiction but quite another to say or imply that all models in science are such. Frigg appears to be arguing for the second while Suárez (2009a, 2009b) has clearly been arguing for the first thesis. However, there are ambiguities in these two theses: it is not clear how much Frigg is committed to the idea that models are full-fledged fiction; he seems to be saying rather that there are many significant similarities between models and fiction so much so that we should take the parallel seriously if only for an understanding of some of the so far neglected aspects of models. While it is possible to concur with Suárez and then simply conclude that not everything in science can or should be taken seriously (scientists sometimes do wrought fictions!), that is clearly not what Suárez intended. Such fully fictional models as the model of quantum measurement (Suárez 2009b) should, according to Suarez, be taken very seriously. The conclusion that Suárez draws comes in fact not too far from Frigg’s conclusion, namely, models are fiction and very useful fiction; they are in Suárez’s term “rules of inference.” One question that kept popping up in my mind when I was going through the literature is: why do we need to identify models in science with fiction (in this or that or some other senses) in order to know what they are? Is it sufficient rather to regard them as products of imagination or, simply, hypothetical? I shall keep this question alive in the rest of this discussion.

Since we have already seen arguments to the effect that models are similar to fictional characters and objects, it might be interesting and illuminating to follow through the opposite route: we begin by noticing the dissimilarities between the two, and then bring in the responses of Frigg, Suárez, et al (interpreted versions of such of course) in the hope of obtaining a deeper understanding of the nature of models this way.

1. While models are conceived to represent real stuff in nature, fictional characters are not conceived or intended that way at all. There does not seem to be any serious sense in which models such as the Newtonian model of the solar system or Bohr’s model of hydrogen atom belong to the same category as Sherlock Holmes or Santa Claus.

It is not entirely clear how this challenge is met in the existing literature. One strategy seems to be admitting that models are regarded and intended by their makers and consumers differently as fictional characters are by their makers and consumers; and yet, it is argued, that they are so similar in most other important aspects that they should be identified despite this difference. A seemingly more effective response may point out the fact that all scientific models are so idealized that no systems in reality exist the way as depicted by the models. There are, e.g., no Newtonian solar system or Bohr’s atom in reality, just as there are no Sherlock Holmes or Santa Claus in reality. However, this is beside the point: models and fictional characters are dissimilar even in this very aspect, namely, what makes

Bohr's model 'unreal' is the use of idealization (which is essential to theory construction in science in general), whereas idealization in any shape or form cannot be involved in the creation of fictional characters, or at least good characters are the ones that are as concrete, as non-idealized, as possible; only the bad ones, the ones that we usually see in crude propaganda fiction, are idealized. If one thinks carefully about the parallel invoked here, one can hardly miss the methodological *opposition* of modeling and character creation in literature. The one has a particular – a real system – in the world that it is used to describe and study, and so judicious abstraction is both justified and highly prized. The other has no particular in the world and yet one 'particular' is invented by imagination (not a real particular in the actual world of course but a fully concrete creature in our imagination), and so abstraction is both superfluous and condemned.<sup>11</sup>

One plausible response from the positive view of models and fiction is suggested by the fact that when models are regarded as fictional they are often models of, say, quantum systems, which implies that they are models of the unobservable (see Suárez 2009b). Since we are in general not able to point to a quantum system and say that it is that of which our model is an account in abstraction, our situation with such models are similar to those with fictional characters. Again, this seems to be a misplaced point. Fictional characters are not just unobservable, they are non-existent by default. No quantum systems – not their models – can be assumed by default in this way; in fact the right metaphysical attitude towards such systems should arguably be that they are just like the macroscopic, observable systems, namely, actual particulars, which happened to be unobservable (a condition that we, the observers, have); and with such an attitude, we imagine in the process of modeling abstract systems that we think could describe the unobservable systems in the same way models of observable systems do. Some scientists deviate from this attitude and begin to treat the models as if they are the real thing, then they are indeed treating them as fiction and they are wrong; they have confused the representation with the represented and part of philosophers' job is to correct them. It would be ill-advised to take such scientists as having obtained some important philosophical insight and draw philosophical conclusions from that, such as thinking that models are fiction after all.

I do think this last response has some truth in it, and yet this is far from being correct simpliciter. To get to what really is the case on this matter, we need to see other aspects of this debate.

2. It appears that what models and fictional characters have in common are just that they are mental entities and they are creatures of our imagination. Too many heterogeneous categories of entities share these two features and yet there is no point of identifying them (imagine someone argues for the similarities between

---

<sup>11</sup> We should realize that both models and fictional characters are *abstract* in the minimal sense that they do not exist spatiotemporally. But this minimal condition they share does not negate the obvious differences I am trying to point out here. We could think of the difference between models and fictional characters as the difference of degrees of abstraction. I shall come back to this point later.

models and theories or models and hallucinations; both would be pointless but for opposite reasons).

To this point, Frigg gives us four reasons for identifying models with fiction. First, despite its non-referring status, descriptions of fictional characters and objects are meaningful and do inform us about reality; the same is true of models. Second, fictional characters are assumed to have a ‘full existence’ – in terms of all necessary properties for such an existence – in the fictional world despite limited explicit description; the same is true of models. Therefore, and third, it always makes sense to finding out the ‘missing aspects’ in the descriptions of fictional characters; and the same is true of models. And fourth, it does make sense, especially fictional characters and situations in serious literature, to compare what is said with what actually happens.

Correctly understood, these observations are surely right; but the rub lies in how they may be correctly understood, and that depends on what sort of metaphysics one has in mind when doing the comparison as above. If models and fiction are understood in a mistaken or inappropriate metaphysics, those claims could be regarded as rather off base. Here is a brief look of what the difficulty may be. While fictional characters or events may well have their own ontology, it does not seem reasonable to think that scientific models have their own ontology apart from the ontology of the stuff they are used to represent. While statements about fictional characters draw their meaning from a semantics that is supported by the metaphysics of fiction, it does not seem reasonable to think of the semantics of scientific statements – which of course use models – have a semantics that is different from factual statements, or *does it?*

The metaphysics of fiction has always been a small and neglected sub-area in metaphysics<sup>12</sup>, but it has been catching more attention in recent years, partly because of recent works by Thomasson (1999) and Jubien (2009). According to Thomasson, who argues for a realist view of fictional characters and events, the world in which anything fictional exist is the world that is created by the author/maker of the fictional works and exists with such works as cultural artifacts. And the characters and events are real to those people who are competent in ‘handling fiction,’ who have such capacities as being able to clearly distinguish the fictional world from the actual world and being able to recognize such objects across different texts, stages, and screens. Statements that are *essentially about* such objects draw their meaning by having terms referring to the objects in that world of fiction,<sup>13</sup> and the truth-condition for such statements is therefore fixed accordingly. *Prima facie*, a fictional world is unlike the actual

---

<sup>12</sup> Such names as Bentham, Meinong, Vaihinger, and Frege, Russell, and Kripke come to mind, but for a brief history and literature of this area, see Thomasson 1999 and Frigg 2010.

<sup>13</sup> Non-essential statements include statements about actors who play certain characters and dramatizations of fictional events, etc. Any statement that is about anything in the actual world that is used to dramatize the fictional must be regarded as not about the fictional or non-essentially about the fictional, and its semantics has nothing to do with the semantics for fiction.

one in that they do not exist spatiotemporally and there are presumably as many such worlds as has been separately invented in works of fiction<sup>14</sup>, and whatever is ‘discoverable’ in such a world must be there by stipulation and consistency. One can discover something about Harry Potter that is not written in any of the Potter novels as long as it is implied (in the logical sense of implication) by what is written; otherwise nothing is discoverable; however, if laws of nature are assumed in the creation of a fictional world, physical as well as logical possibility and necessity operate in that world.<sup>15</sup>

Rough and incomplete as it is, this sketch of a realist metaphysics of fiction gives us enough idea, I think, to evaluate the comparison of models and fiction. Frigg’s first point as summarized above is good only if we can embrace the same kind of ontology that fictional worlds receive. But is that plausible? Are we ready to defend the idea that there are as many physical worlds as there are different invented models? This may well be a view that Goodman holds (Goodman 1978), and Frigg may well be a disciple of his, although Goodman’s name is nowhere mentioned in Frigg’s or anyone else’s works on models and fiction.<sup>16</sup> And indeed, if we are willing to embrace a parallel ontology for models in science as containing multiple model worlds, all four observations by Frigg about the similarity between models and fiction make good sense. The remaining question is about how these worlds are related to the actual world, which has always been recognized in the literature of scientific representation as a serious problem and for which Frigg invented a term, ‘t-representation,’ to study it separately from the ‘p-representation,’ which refers to the relationship between a representational device and the model it p-represents (see Frigg 2011b). I shall return to this question below.

---

<sup>14</sup> The identity of such worlds has a theory of its own; but fictional worlds can be shared by multiple works as long as they are intended to share the same world, otherwise even if the worlds are more or less related to the same actual world, they may not have anything in common. However, there are problems such as whether Emma’s India in *Pride and Prejudice* is the same as Holmes’ India and whether the India there is a fictional or an actual place (given that India is never described in any detail in either work).

<sup>15</sup> Thomasson (1999) calls her theory of fiction ‘the artifactual theory,’ and thinks of the fictional worlds or characters as “abstract artifacts – relevantly similar to entities as ordinary as theories, laws, governments, and literary works” (p. xi). In the minimal sense, if one accepts Thomasson’s theory, one must think of models as fiction, unless one thinks that models and theories are categorically different. However, I think the sense in which Thomasson identifies fiction as artifactual is so minimal that it does not relieve us from investigating in what exact sense is a model fictional. On the other hand, Thomasson’s somewhat casual identification of say fiction with theories is still controversial. It is not clear whether one can straightforwardly say that characters in a novel are relevantly similar to chemical elements in a theory.

<sup>16</sup> For lack of space I shall not further discuss this connection with Goodman’s work on ‘worldmaking’ and science in general. It will be the focus of another full-length study.

3. While most of the scientific models refer to types of things, fictional creations are mostly focused on particulars, such as characters and events.<sup>17</sup> When it comes to types of stuff in a fiction, the names usually refer to real stuff in the world. When Conan Doyle describes the gun Sherlock Holmes uses, it is of a type that does exist in the world. The same is true with people or animals or streets and building, namely, when they are thought of as types of things, they are just ordinary types in the world; although the particular person or animal (e.g. the Hound of Basqueville) or building (e.g. the Mansion on the Wuthering Heights) is fully fictional. There are some fully fictional types, such as Spiderman's glue from his palms or quantanium in "Giants and Aliens", etc., but they are minor categories in the realm of fiction. So, models and fictional objects are nothing alike, or so this objection claims.

There is a rather big category of types, such as unicorns, satyrs, etc. if we count mythological creatures, but still they are markedly different from models. The fictional creatures all have normal, realistic enough parts, while scientific models, especially those for unobservable objects, are typically constituted by parts whose existence is the focus of debate. It is in this sense, a model of water is the opposite of unicorn; while water as the whole system being modeled is sensible and real enough but the molecules and atoms which are modeled to compose it are speculative, or even fictional, one may say, parts of unicorn are shared by normal animals only the whole system is fictional. What does this disparity implies? One superficial answer seems to be that models represent while fictional objects don't. The scientific model of water, whether of a whole sample or of its macroscopic parts, is intended to represent whatever stuff that's in the actual world, while the fictional 'model' of a unicorn needs no or little representation of its parts and does not represent anything as a whole.

To summarize the above in the form of an objection to taking models as fiction, one could say that (1) even though one can create a similar metaphysics for models as for fiction, there does not appear to be enough motivation for doing so; (2) models are typically representational devices while fictional characters and events are typically not; (3) models are mostly types while fictional things, at least the important ones, are mostly particulars; and (4) when fictional objects are types, they are mostly other-worldly, which indicates that when the same categories of things are compared between models and fiction, they really don't have anything in common (other than the minimal similarity, namely, they are mental or cultural/communal rather than material).

---

<sup>17</sup> Note here, we think of fictional characters and events as particulars only in the fictional world. Sherlock Holmes is an individual in the Holmes' world because he does occupy space and time in that world, although he does occupy any actual space-time regions (some books in which he 'lives' do occupy such regions but that's beside the point). Also, electromagnetic force line in a model cannot occupy any space-time regions, actual or otherwise; only its tokens can do that. And so, it is not even a particular in any imagined worlds.



## 6 Conclusion

Here is what I think is the case between models and fiction, the arguments for which, too long to be given in this paper, are given in another place. Duhem is *right and profoundly so* when he distinguishes the A method and its product from the C method and its. Philosophers today are wrong when they think of products of the A and the C method as belonging to the same category of models, only that one is more abstract than the other. We can, and people do, call all of them models: the point-line image of the solar system is a model of it, and the Bohr's model of hydrogen atom is also a model, and yet if the above analysis about models and fiction is right, these two types of 'entities' are of radically different nature. The first kind, results of the A method, are representational devices, while the second kind, results of the C method, are not. Roughly speaking, the A-models are used for observable systems, while the C-models are used for unobservable ones. They are radically different because they have fundamentally different origins. The A-models, the abstract models of observable systems, are results of idealization, whose purpose is purely practical and they are dispensable once the mathematical theory is fully mastered. The C-models, the abstract models of unobservable systems, have their origin, I now suggest, deep in the tradition of mythology.<sup>18</sup> In this sense, the A-models have nothing to do with fiction, while the C-models are fully fictional, as mythology and the like are fictional. They are not created for practical purposes, such as saving intellectual effort, as the A-models do; they are created to fill a deep-seated need of us humans to know what is going on behind/beneath the observable phenomena. It is in the sense that scientists who investigate the microscopic world have been creating worlds in which postulated creatures are responsible for the few glimpses we humans gain through our limited means of experiments that we see the origin of the C-models. And in this sense, they are fully fictional. This explains why the models for the observables are all very abstract while those for the unobservables are often deliberately endowed with observable qualities (e.g. the 'tangibility' of force lines the visual 'realness' of a pudding model for atoms). The former is obvious because why would we want to have a qualitatively similar models except perhaps for educational purposes, while the latter is so because the models are all we have so long as we want to know what the systems we are after may look like. There are assumed abstractions in such models but the abstractness is a matter of assumption given the scarcity of evidence; it is definitely not the result of idealization since there is nothing for us to idealize about. From what we may ask is the identicalness of the elementary particles an idealization of? And because of this fact, many philosophers of science (starting from van Fraassen 1980, and more recently French 2006) think that elementary particles are identical! Can we imagine anybody argues for the idea that because most of the time a mass point is what the earth is in our model that earth is really a

---

<sup>18</sup> Here the term 'mythology' is used in its broadest sense. It might be too much to call Plato's philosophy a mythology, but certainly I am regarded it here as having heavy mythological elements.

dimensionless mass object? Because of observability (or the lack of it) of the whole systems in question, what is a type or a particular is treated differently in the A-models and the C-models.

I mentioned earlier a possible philosopher's objection to some scientists' tendency of reifying the models of the unobservables. The objection takes such an attitude towards the models of the unobservables as mistaken because we do not seem to have any good reason not to think that the unobservables are just like the observables as they are in themselves, only they are too small or too far away for us to observe. If so, we have no reason then to think of models for unobservables as anything more than representational devices, and what I have just said couldn't possibly be right. I also said earlier that this may not be the final word for the matter. Given the origin of such models, and given how long and hard people have struggled to construct the right models for the unobservable world, I would argue that the fictional view of models has more support from the history of science and from philosophical considerations. If Locke's admonition is right about getting to know what the secret and unobservable powers of the external world are, which produce the impressions of color, sound, and texture in our perception, I would argue that this fictional view of the models about those secret objects and their powers draws some support from Locke's theory of perception.

Russell, as I discussed earlier, is surely right when he says that only structural, i.e., relational, properties are knowable via science, and qualitative properties are only knowable by acquaintance. If so, the qualitatively described model-systems about things we have no acquaintance of can only be fictional systems.

And finally, Frigg is not right in that not all models are representational, nor are they all fictional. Suárez is right that some models are fictional, but he is not when he thinks of such models as only of inconsistent systems or states of affairs.

But the ultimate question still remains for the C-models. If they are not really representations as the A-models are, how do we explain the fact that they are thought of, and used as, representations in science? Can the scientists and laymen alike who hold this attitude be completely mistaken? Well, no, of course they are not entirely mistaken; there is a reality beyond the observables that is causally responsible for the observables, and the C-models are undoubtedly about that reality. However, they are about that reality, call it representation if you like, not in the same way that the A-models represent, namely, a idealized construction of observable systems for practical convenience. And this is precisely why I think C-models are fully fictional. Fictional characters and events are also about reality, about the actual people and events; but they do not represent them *per se*, or they do not represent them the same way figures and events described in a history or biography are used to represent. These latter are somewhat similar to the A-models (though they are quite different in many other ways) because they are idealized descriptions of the real things. Fictional characters and events, at least those that are created by the most esteemed minds in the history of literature, are about a deeper and 'more real' reality. They tell us about humanity in a way that we cannot get from reading histories and biographies. In this respect, the C-models are exactly like fictional objects. They show us a reality that we never

learn from putting together the observable results of however exhaustive studies of reality. Think of Mach's suggestion for science (Mach 1950. 1984), which goes roughly as this: science should be no more than the most economic organization of the observed and observable results. One won't find a place for the C-models in that kind of science and therefore one won't find the kind of understanding of reality that the C-models provides. So, Mach and to a great extent Duhem could be interpreted as great champions for giving no place to 'fictional thinking' in science. They have grudge against the A-models because they are intellectually economic and non-speculative, while C-models are neither, so they should be given a place in science.

Finally, we have noted earlier, especially in our discussion of Hesse's view on modeling, that many models are created by analogy and metaphor. It's obvious that no A-models are created that way; metaphorical reasoning does not apply to the creation of a point-line model for the solar system. However, when Bohr thought that atoms are like solar systems, metaphorical reasoning was in full swing, and Bohr's models for atoms are typical C-models. This is also what happens in fiction, or so shall I argue. We often say that a fictional character has a 'real-life model'; which is sense of 'model' that we haven't seen in the literature of scientific representation. Let's suppose that Conan Doyle was the real-life model for Watson, we want to know whether there is anything going on in science that resembles this common phenomenon in literary works? Take Bohr's model for hydrogen atom. It's 'model' on the Newtonian mechanical model for the solar system. It appears that there is, *mutatis mutandis*, a significant difference between these two cases. The model (in this particular sense of 'model') for the Watson character is Conan Doyle, a real person, while the model for Bohr's model is the Newtonian model for the solar system, another model. But wait, to say that Conan Doyle, the real person, serves the model for the Watson character, is just an elliptical way mentioning the phenomenon of literary creation. Who can use a real person or object as the model? What must be the collection of impressions or understandings of the person or object that serve as the 'model' – meaning the basis – for the imaginary creation of the character. Understood this way, similarity between the two cases is next to perfection. And this is certain the ways in which our ancestors created gods and other deities, for even though they have models in mortal human beings, they refer to creatures who are certainly not; and this is why I say that the origin of such models as Bohr's model of hydrogen atom is not anything remotely like the modeling of the solar system. It rather lies in the deep past of myth-making.

## References

- Bokulich, A.: Open or Closed? Dirac, Heisenberg, and the Relation between Classical and Quantum Mechanics. *Studies in History and Philosophy of Modern Physics* 35, 377–396 (2004)
- Callender, C., Cohen, J.: There Is No Special Problem About Scientific Representation. *Theoria* 55, 7–25 (2006)

- Daitz, E.: The Picture Theory of Meaning. In: Flew, A. (ed.) *Essays in Conceptual Analysis*, pp. 53–74. MacMillan, London (1956)
- Demopoulos, W., Friedman, M.: The Concept of Structure in The Analysis of Matter. In: Savage, C.W., Anderson, C.A. (eds.) *Rereading Russell*. *Minnesota Studies in the Philosophy of Science XII*, pp. 183–199. University of Minnesota Press, Minneapolis (1989)
- Demopoulos, W.: Russell's Structuralism and the Absolute Description of the World. In: Griffin, N. (ed.) *The Cambridge Companion to Bertrand Russell*, pp. 392–419. Cambridge University Press, Cambridge (2003a)
- Demopoulos, W.: On the Rational Reconstruction of our Theoretical Knowledge. *Brit. J. Phil. Sci.* 54, 371–403 (2003b)
- Duhem, P.: *The Aim and Structure of Physical Theory*. Princeton University Press, Princeton (1954)
- Freeman, W.J.: The Physiology of Perception. *Scientific American* 264(2), 78–85 (1991)
- French, S.: Identity and Individuality in Quantum Theory, *Stanford Encyclopedia of Philosophy* (1989), <http://plato.stanford.edu/entries/qt-idind/>
- French, S.: Structure as a Weapon of the Realist. *Proceedings of the Aristotelian Society* 106, 109–187 (2006)
- Frigg, R.: Models and Fiction. *Synthese* 172, 251–268 (2010)
- Frigg, R.: Fiction in Science. In: Woods, J. (ed.) *Fictions and Models: New Essays*, Philosophia Verlag, Munich (2011a) (forthcoming)
- Frigg, R.: Fiction and Scientific Representation (2011b) (preprint)
- Goodman, N.: *Ways of Worldmaking*. Hackett Publishing Company, Indianapolis (1978)
- Hesse, M.G.: *Models and Analogies in Science*. University of Notre Dame Press, Notre Dame (1966)
- Hughes, R.I.G.: Models and Representation. *Philosophy of Science* 64 (Proceedings), 325–336 (1997)
- Hylton, P.: *Russell, Idealism, and the Emergence of Analytic Philosophy*. Clarendon Press, Oxford (1990)
- Jubien, M.: *Possibility*. Oxford University Press, Oxford (2009)
- Mach, E.: *The Science of Mechanics: A critical and Historical Account of its Development*. McCormack, T.J.(trans.) Open Court, La Salle (1960)
- Mach, E.: *The Analysis of Sensations and the Relation of the Physical to the Psychical*. Williams, C. M.(trans.) Open Court, La Salle (1984)
- Morgan, M.S., Morrison, M. (eds.): *Models as Mediators*. Cambridge University Press, Cambridge (1999)
- Russell, B.: *My Philosophical Development*. George Allen & Unwin, London (1959)
- Russell, B.: *The Analysis of Matter*. Routledge, London (1927/1992) (originally published in 1927)
- Salmon, W.: *Scientific Explanation and the Causal Structures of the World*. Princeton University Press, Princeton (1984)
- Siegel, S.: *The Contents of Visual Experience*. Oxford University Press, Oxford (2011)
- Strawson, P.: *Individuals: An Essay in Descriptive Metaphysics*. Methuen, London (1959)
- Suárez, M.: Scientific Representation: Against Similarity and Isomorphism. *International Studies in the Philosophy of Science* 17, 226–244 (2003)
- Suárez, M.: Fictions in Scientific Practice. In: *Fictions in Science: Philosophical Essays on Modeling and Idealisation*, pp. 1–15. Routledge (2009a)

- Suárez, M.: Scientific Fictions as Rules of Inference. In: Suárez, M. (ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization*, pp. 158–178. Routledge (2009b)
- Teller, P.: Twilight of the Perfect Model Model. *Erkenntnis* 55, 393–415 (2001)
- Thomasson, A.L.: *Fiction and Metaphysics*. Cambridge University Press, Cambridge (1999)
- van Fraassen, B.: *The Scientific Image*. Clarendon Press, Oxford (1980)
- van Fraassen, B.: *Scientific Representation: Paradoxes of Perspective*. Clarendon Press, Oxford (2008)
- Walton, K.L.: *Mimesis as Make-Believe*. Harvard University Press, Cambridge (1990)
- Wittgenstein, L.: *Tractatus Logico-Philosophicus*. Pears, D. F., McGuinness, B. F.(trans.), Russell, B.(intro.) The Humanities Press, New York (1963)

# From the Received View to the Model-Theoretic Approach

Leilei Qi and Huaxia Zhang\*

**Abstract.** The nature and structure of scientific theories has long been one of the cores of philosophy of science. Since the failure of the Received View of the logic empiricists', the concept of "paradigm" of the historical school paid attention only to the structure of scientific revolution while ignoring the structure of scientific theories. While the new empiricism studies the analogy model, it lacks precise and systematic analysis of scientific models. In recent years, philosophers inclining to logic and analytic philosophy and not satisfied with the historical approach have attempted unceasingly to find a new approach. They've found a way in the state space theory especially in set theory, taking models as the core of scientific theories and set theory as its semantic analytical tool. This trend has gradually entered the analysis of theoretical structure in philosophy of science, forming the model-theoretic approach of scientific theories.

The Received View in philosophy of science is the logical empiricists' way of analyzing the structure of scientific theories. It is the central theme of philosophy of science for quite a long period in the 20<sup>th</sup> century, but encountered enormous difficulties afterwards so that one of its founders Carl Hempel announced publicly he had to give it up. Along with the death of several philosophers of science, i.e. Thomas Kuhn, Paul Feyerabend and Carl G. Hempel in 1996-1997, the central

---

Leilei Qi

Research Center for Philosophy of Science and Technology,  
South China University of Technology, Guangzhou, P.R. China  
e-mail: pollqi@scut.edu.cn

Huaxia Zhang

Department of Philosophy, Sun Yat-sen University, Guangzhou, P.R. China  
e-mail: hsszhx@mail.sysu.edu.cn

\* Supported by the Fundamental Research Funds for the Central Universities, SCUT (2011SM023).

thesis of philosophy of science experienced a to-and-fro process from abstract theory to experimental experience then to abstract theory. In recent years modern structuralists have built up the banner of the structure of scientific theories through set theory and model theory. This research has become a new trend. Approaches from the Received View to the model-theoretic approach are competing and complementing each other and thus greatly enrich the research content of the structure of scientific theories.

## 1 To Give Up the Received View

Having systematically integrated John S. Mill's inductionism, Ernst Mach's positivism, Gottlob Frege and Bertrand Russell's logical theory, logical positivists aimed to find the logical structure of scientific theories and put forward the full view of scientific theories. The two leading figures of logical empiricism Rudolf Carnap and Carl G. Hempel made separate analysis and interpretations of scientific theories, which were widely supported before 1960s. This is commonly called *The Received View* or *the Standard Account of scientific Theories*.

The Received View is a statement or syntactical view on scientific theories. That it takes scientific theories as linguistic entities characterized by syntactical features, and a scientific theory typically consists of an axiomatic theory system and a set of correspondence rules. The axiom system is a set of theoretical laws formulated by first-order logic language  $L$ , and the so-called theoretical law is a general statement containing the *theoretical vocabulary*  $V_T$  which plays a fundamental axiom role in the axiom system. The set  $C$  of correspondence rules is a set of interpretative rules connecting theoretical vocabulary  $V_T$  and observation vocabulary  $V_o$ , and provides experiential meaning through observation vocabulary. Its logical form is:

$$(x)(F_x \equiv O_x)$$

Here ' $F_x$ ' consists of  $V_T$ . ' $O_x$ ' is an expression of  $L$  containing symbols only from  $V_o$  and possibly the logical vocabulary. In this way, the so-called theory is TC.

The Received View is the epistemic heart of logical positivism. The dogma of logical empiricism may be understood as based on the Received View such as follows:

### (1) Rejecting metaphysics

Since the Received View admits theoretical entities, other metaphysical entities as "vitality" and "ether" must be given an explicit definition in terms of the observation vocabulary by correspondence rules. Those that can not be given such explicit definitions should be excluded from science. It works like a firewall, metaphysics is separated out and discarded.

### (2) Positivism

It means the meanings of vocabularies are based upon their proven methods. It is required to find the proven methods of theories and the theoretical vocabularies by correspondence rules.

### (3) Inductive logic

How can one prove a scientific theory by observation? By establishing a set of inductive logic.

Although logical empiricists had been improving their theories, its limitations had not been explored. On March 26, 1969, opening the Illinois symposium on the Structure of Scientific Theories, with 1, 200 persons in the audience, Carl Hempel, as the keynote speaker who was expected to present the latest revision of the Received View, explained why he was abandoning both the Received View and reliance on syntactic axiomatization [1]. He said, “The terms of the antecedent vocabulary are by no means assumed to be ‘observational’ in the familiar theoretical-observational distinction, ...they are not required to stand for entities or characteristics whose presence can be ascertained by direct observation unaided by instruments and theoretical inference” [2, p.245]. He also said, “I turn now to another difficulty of the standard view” to that schema TC, “I have come to feel increasing doubts about its adequacy...Some brief remarks on the concept of correspondence rules as constituents of a scientific theory. The customary designation of the sentences in question as ‘rules’, or as coordinative or operational ‘definitions’, strongly conveys the suggestion that they constitute truths guaranteed by terminological legislation or convention. But this idea is untenable for several reasons” [2, p.252]. As to the role of model in theories, the standard view reckons that “analogical models can be of considerable didactic and heuristic value” only, but “it seems to me to play an essential role in the formulation and application of many theories” [2, pp.251-252].

Gradually, philosophers inclining to logic and analytic philosophy and not satisfied with the historical approach have attempted to find a new approach to replace the logical empiricists' Received View and the historical school. They found a way in set theory and model theory, regarding models as the core of scientific theories. This research trend gradually entered the analysis of theoretical structure in philosophy of science, forming the model-theoretic approach of scientific theories. It is noteworthy that the British representative of the Received View, A. J. Ayer, said in an interview, “I suppose the most important [defect]...was that nearly all of it was false” [3]. However we believe that the approach to study scientific theory through the axiomatic method is not wrong. The reason is that it can typically sort out the basic concepts, composition and structure of theories, make comparisons between theories, study theory reduction and emergence, explore the unity and diversity of science, and analyze these problems through mathematical methods. The problem is that logical empiricists merely linguistically or syntactically analyzed such formalizations of scientific theories, while the alternative method is the semantic or model-theoretic approach.

## 2 The Model-Theoretic View of the Structure of Scientific Theories

The Model-Theoretic view of the structure of scientific theories may be traced back to the American scientist John Von Neumann in the late 1930s and the



mathematician George David Birkhoff in the late 1940s. It may be considered as the real source of model theory that the Dutch philosopher and logician Evert W. Beth studied logic, analyzed several specific theories through set theory, inferred the potential of semantic analysis, and advocated amplified semantics. Beth's semantic tableau is a proof method for formal systems. It is considered by many people, especially students not acquainted with logic, to be intuitively simple. The original model theory, therefore, was developed as a branch of mathematical logic.

Some philosophers began to research the theoretical structure through the model view of scientific theories the 1970s. Representatives include F. Suppe, Bas c. Van Fraassen, P. Suppes and J.D. Sneed. Suppes is the first to systematically develop the theory. Bas c. Van Fraassen and F. Suppe used to the non-formal notions different from but based on the logic axiomatic system, which is a state-space semantic model-theoretic approach. Suppes' student J D. Sneed integrated Suppes' observational-theoretical term distinction and E.W. Adams's results to build a very sophisticated axiomatic system on scientific theories. W. Stegmüller and others developed J D. Sneed's theory. The book *An architectonic for science: the structuralist program (1987)* co-written by W. Balzer, C.U. Moulines and J D. Sneed marked the maturation of the theory, and eventually formed the outlook of scientific theories of structuralism.

The model-theoretic view of the structure of scientific theories is the opposite of the Received View of logical empiricism. It states that a theory is neither a statement collection nor a linguistic entity, but an extralinguistic entity described by set theory. Its basic feature is to deny that a theory is a statement, while arguing that scientific theories are axiomatized in set theory through defining predicates of a set theory. Therefore, if a theory is axiomatized by predicates of a set theory, anything meeting that definition is a model of the theory.

From the point of view of set theory, a structure  $M$  is given by the following factors : the base set  $D$  written as  $\text{dom}(D)$ , the relation set  $R$  on  $\text{dom}(D)$ , the function  $F$  from  $\text{dom}(D)$  to  $\text{Dom}(D)$ , and the individual constant  $c$  on  $\text{dom}(D)$ . Thus structure  $M$  is a non-linguistic entity, i.e.  $M = \langle D, R, F, c \rangle$ . When formal language  $L$  is attached to describe the structure,  $M$  can make interpretation of all the symbols of  $L$ . For example, the structure of real number interprets the symbols of  $-$ ,  $+$ ,  $\times$  in real number theory as "negative", "adding" and "multiplication". Some of the languages and formulas in  $L$  have their true value (true or false) in the non-linguistic structure  $M$ . If a set of sentences or formula  $\Psi$  of  $L$  is true in  $M$ ,  $M$  may be taken as a model of  $\Psi$ , which is:

$$M \models \Psi.$$

It also means that  $M$  satisfies (or interprets)  $\Psi$ . If  $T$  is a theory expressed by  $L$ ,  $M$  is called a model of  $T$  when all valid sentences  $\Psi$  of  $T$  are true in  $M$ . It meets the famous saying of the logician Tarski, "a possible realization in which all valid sentences of a theory  $T$  are satisfied is called a model of  $T$ " [4, p.11]. This is a common definition of model to all the model-theoretical approaches in the research of scientific structure.

Specifically, the semantic model described by set theory believes that  $X = \langle D, R \rangle$  is a model, which is that “X is S”, where “S” represents any scientific theory, and “is S” is the predicates of set theory. The prerequisite that “X” meets “is S” is that the contents of the ordered pair  $\langle D, R \rangle$  must be defined by several specific conditions of D and R, and these specific conditions are just to axiomatize the laws of scientific theories, or to say to define S through the definition of D and R. It may be further understood in this way that it is all theorems of axiomatized scientific theories that describe each element of a model as well as its relationships with other elements.

Researchers with different interests prefer different expressions for this research approach, such as the “non-statement view”, the “semantic approach”, the “model approach”, the “set theory approach”, “structuralism” and so on. For example, F. Suppe based his approach on semantics, Suppes’ (1969) based his upon set theory, and Van Fraassen’s (1972, 1980) new scientific picture “state-space approach” [4, p.67]. Among those new theories, Sneed (1971), Balzor and Monlins (1996) emphasized Suppes’ role, and are now called the Stanford School in philosophy of science. Others who emphasized Sneed’s role are the called Sneedean School. Van Fraassen clearly pointed out that there were two different lines of research. He said, “With respect to the structure of physical theory I see two main lines of approach: one deriving from Tarski and brought to maturity by Suppes and his collaborators (the set-theoretic structure approach) and the other initiated by Weyl and developed by Evert Beth (the state-space approach)..... My own inclination in that subject area has been toward the state-space approach.” [5, p.67] The paper is to introduce these several philosophers’ theories in the following sections.

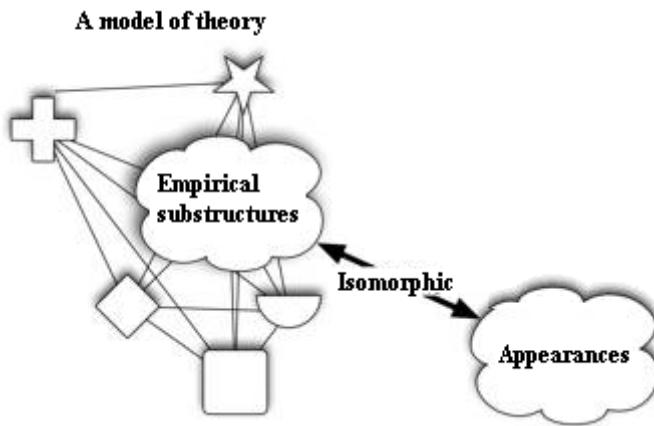
### 3 F. Suppe’s and C. Van Fraassen’s State-Space Model

Suppe said that the “theories are not collections of propositions or statements, but rather are extra-linguistic entities which may be described or characterized by a number of different linguistic formulations” [6]. This implies that examining linguistic formulation is not enough for understanding theories. The contents and structure of scientific theories may be understood through the semantic or model-theoretic approach. It seems that F. Suppe and C. Van Fraassen are the first to take this approach in a systematic way.

The same theories usually have various linguistic formulations, for example, wave mechanics and matrix mechanics are equivalent formulations of quantum theory, so it is a necessary to centre on “the semantic concept of theory” to examine their structures. Suppe summarized his idea of the semantic model approach in the following way: (1) A scientific theory is a theoretical structure in the sense of set theory. It is an entity of set theory whose domain is selected by scientists to answer a variety of questions. (2) The structure of a theory explains the behavior of a phenomenon system through the medium of physical systems. (3) A physical system is the replica of a phenomenon system. It does not attempt to describe all aspects of the phenomenon in its domains, but to abstract certain parameters that

exert an influence on the phenomena. Such idealized replicas of phenomena are called physical systems. (4) A particular configuration of a physical system is the state of an entity, so that a set of values for the parameter is a representation of the state. Therefore, the state of a physical system is represented as n-tuples of numbers, and the theoretic laws of physical systems represent the changes of states over times. [6, p.223]

Van Fraassen proposed a new picture to understand theories. He argued that scientific theories have three aspects: (1) To present a theory is to specify a family of structures, namely its models, which satisfy a set of axioms or propositions of the theory. (2) To specify certain parts of those models as the empirical substructures for the direct representation of observable phenomena. The structures which can be described in experimental and measurement reports are called appearances. (3) The relation between theory and phenomena is not true or false but instead is judged by empirical adequacy. The theory is empirically adequate if it has some model such that all appearances are isomorphic to empirical substructures of that model [5, pp.64-65]. (4) A physical theory then typically uses a mathematical model to represent the behavior of a certain kind of physical system. A physical system is conceived of as capable of a certain set of states, and these states are represented by elements of a certain mathematical space, the state space. Specific examples are the use of Euclidean 2n-space as phase space in classical mechanics and Hilbert space in quantum mechanics[7]. The following figure 1[8] may describe the empirically adequate model:



**Fig. 1**

In summary, Suppe emphasizes the abstract and idealized replicas of science and Van Fraassen emphasizes the model of state-space.

Their common ideas are as follows:

(1) A scientific theory has its subject matter that is a class of phenomena known as the intended scope of the theory. It does not attempt to describe all the aspects of the phenomena.

(2) The first step of scientific theory is to abstract a few set of parameters of entities from the intended scope of phenomena with the isolated condition under experimental control. The next step is to introduce the ideal or even fictional state and conditions for the parameters to act. The last step is to construct physical systems as theoretical models to express how the parameters change as the states change and vice versa. One parameter is expressed in one dimension coordinates, and n-tuples parameters are expressed by n-dimension coordinates, which form n-dimensional phase space. To give the simplest example, a classical particle has, at each instant, a certain position that needs three dimensional spaces, namely  $q = (q_x, q_y, q_z)$ , its momentum also requires three dimensional spaces, namely  $P = (p_x, p_y, p_z)$ , and thus its state space may be Euclidean 6-space, whose points are the 6 tuples of real numbers  $(q_x, q_y, q_z, p_x, p_y, p_z)$ . And to describe the motion of two particles requires 12-dimensional spaces. The motion and change of physical entities represent themselves as the behavioral trajectory. Here the physical system does not mean the phenomenon system but the system of physical parameters as well as their relations in the phase space. That is the semantics and ontological commitments of theories.

(3) There is a question about the relationship between models and reality since the parameters of a model physical system can be measured in the phenomenon world (real world). Only when the data from measurement are translated into a special form can they be compared with the results predicted by the model physical system to check whether the model predicts and is confirmed by the phenomenon world. Suppe supposed a triple relation between the theoretic proposition, the model-semantic contents of a theory and the real world, using and reconstructing realism-instrumentalism disputes as in the following figure 2:

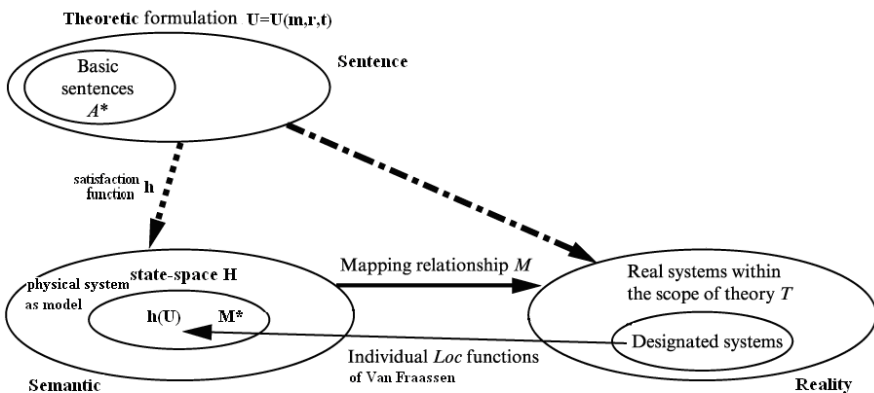


Fig. 2

Here theoretical structure  $T$  is composed semantically by the intended interpretations of formulation  $U$ . The formulation  $U$  is not divided into observational and theoretical and so called corresponding rules. Formulation languages  $U$  are also interpreted as referring to real systems within  $T$ 's scope. Theories provide the state-space interpretation of physic system and also are asserted to stand in some mapping relationship  $M$  to real systems within the scope of the theory. On a realistic version,  $M$  would be a homomorphism; on Suppe's quasi-realistic version,  $M$  would be a counterfactual relationship specifying how the real systems would behave were they isolated from influence by variables not in  $T$ ; on van Fraassen's constructive empiricism the mapping  $M$  is between a designated subset of the real systems and its image  $M^*$  under individual *Loc* functions. When  $M^*$  is contained in  $T$  then  $T$  is said to be empirically adequate. The *Loc* functions specify the ontological commitments one makes in asserting the theory  $T$ . On Suppes' version, the mapping relation  $M$  is mediated by a hierarchy of models including models of the experiment and models of data. Experimental design, instrumentation and so on, are not proper parts of theories but are used to determine whether mapping  $M$  does hold[9].

#### 4 Suppes' Semantic Model by Using Set Theory

Unlike the state-space model, the main modeling tool of Suppes is the non-linguistic structure with different kinds of set. His slogan is that "to axiomatize a theory is to define a set-theoretical predicate"[10]. The so-called set-theoretical predicate is something like "X is a group", "X is a decision theory", "X is classic mechanism", and if they can all be expressed by set-theory languages, they are all set-theoretical predicates. The above method to reconstruct and axiomatize a theory is in fact the approach of Bourbaki in mathematics. For Bourbaki, to axiomatize a mathematical theory is no more or less than to define a kind or species of structure in set-theoretic terms. Therefore, Suppes' model-theoretic approach in philosophy of science is the same with the species of structure of Bourbaki. It is composed of the following four points:

(1) A certain number of sets  $E_1, \dots, \dots E_n$

They are the principal base sets to constitute a theory. They are the main basic material to constitute the species of the theoretical structure  $\Sigma$ , like bricks for a building.

(2) A certain number of sets  $A_1, \dots, A_m$  in the theory

They are the Auxiliary base sets to construct the species of the theoretical structure  $\Sigma$  such as the set of real numbers, the set of natural numbers and so on.  $\Sigma$  possibly contains no auxiliary base sets, but it must contain at least one principle base set.

(3) A typification  $T(E, s) = s \in S(E_1, \dots E_n, A_1, \dots A_m)$ , and  $E = \{E_1, \dots E_n\}$

Here "S" is an echelon construction scheme based on the above "n + m" terms of principal base sets and Auxiliary base sets, and "s" is a series of inductive procedures to construct  $\Sigma$ . Each step is composed of the Cartesian product ( $E \times F$ ) of

two sets obtained in former steps or the power set of  $E$ , for example,  $P_0(E)$ , (here  $P_0$  means power set) and the last step obtains the echelon construction scheme.  $T(E, s)$  is called the typical characterization of species of structure  $\Sigma$ , and it further gives the schema “S” canonical extension.

(4) A relation  $R(E, s)$

It is transportable (in theory) with respect to the typification  $T$ .  $R$  is called the axiom of the species of structure  $\Sigma$ .

Cartesian product sets and the power sets are so enriched and their arrangements and combinations are so multiform that they can express all mathematics. As set theorists usually say, the language of set theory is a kind of universal language with which all mathematics (and practically all of our scientific thinking as well) may be reproduced. That is why the semantic approach is so useful and therefore important. If theories may be axiomatized in this manner, we’ll have the whole of mathematics and sciences “at hand”. That reminds Zhang (one of the authors of this paper) of his father. When his father was studying differential and integral calculus in Guangdong Normal School (predecessor of Sun Yat-sen University and South China Normal University), the textbook was written by P.F. Smith of Yale University published in the late 19<sup>th</sup> century. However, in the 1960s, when Zhang was teaching Calculus, he found a textbook from USSR. Although the author was some other person, the content was fully the same as the book of Smith but with only one chapter added before the theory of limits to talk about set theory and the set of real numbers.

Based on set theory, the calculus can well explain various natural laws and economic laws expressed through differential equations. As for discrete mathematics, it is an undisputed fact that it based on set theory. In his famous book *Philosophy of physics*, R. Torretti said, “Imagine for a moment that a demon of uncommon intelligence undertook to do physics in this style. The set-theoretic hierarchy is so rich that he or she could well build from it a species of structure of which nature in all its complexity is an instance.”[11] Of course, axiomatization has its limits, but the model of set theory is still a useful tool to analyze the structure of scientific theories in meta-science.

## 5 The Model-Theoretic Approach of the Sneedean School in Philosophy of Science

The Sneedean School studies model semantics most systematically and thoroughly. J. D. Sneed and his proponents classified science into various models and made analyses of them, such as potential models, actual models, partial potential models, models constraining and linking other models, blurring models and so on. Those classes of models, connecting with experimental models and data models researched in detail by P. C. Suppes, form the whole system of model-class in meta-science. The structures of and relationship between these classes of models embody the structure of scientific theories and model-based reasoning in science.

The following elucidates these classes of models through the example of classic collision mechanics (CCM).

“ $M_p$ ” denotes a class of potential models. It is the conceptual framework of the theory. (In CCM, the potential model contains a set of billiard balls as particles, a set of times  $T=\{t_1, t_2\}$ , the mass of the billiard balls, as well as the positions and velocities.)

“ $M$ ” denotes a class of actual models. It is the subclass of potential models satisfying the empirical laws of theories. (It is the conservation law of momentum in CCM.)

“ $M_{pp}$ ” is a class of partial potential models. It is the non-theoretical basis of theories. (In CCM it only contains position and velocity as a function of time.)

“ $C$ ” is a class of constraints. It is the conditions connecting different models of the same theory. (In CCM it is the conservation of mass and the sum rule of mass.)

“ $L$ ” denotes the class of links. It is the conditions connecting models of different theories. (In CCM, for example, it is the links to classic mechanics, kinematics and relativistic collision mechanics.)

“ $A$ ” denotes a class of admissible blurs (degrees of approximation admitted between different models in CCM.)

As a result, to axiomatize CCM means to express CCM by model theory or set theory, thus:

CCM is classic collision mechanics, if and only if there are  $t_1, t_2$ , such that

- (1)  $P$  is a finite non-empty set.
- (2)  $T$  contains exactly two elements, namely  $T = \{t_1, t_2\}$ .
- (3)  $v: P \times T \rightarrow \mathbb{R}^3$ . Here  $\mathbb{R}$  means real number in three dimensions.
- (4)  $m: P \rightarrow \mathbb{R}^+$
- (5)  $\sum_{p \in P} m(P) \cdot V(P, t_1) = \sum_{p \in P} m(P) \cdot V(P, t_2)$ .

So, through axiomatizing a theory in set theory, the composition of science may be analyzed and cognized. A theory is composed of different classes of models. Its core is  $K$ , and  $K = \langle M_p, M, M_{pp}, C, L, A \rangle$ . Theory  $T$  is equivalent to the ordered pair of  $K$  and  $I$ , namely  $T = \langle K, I \rangle$ . Here “ $I$ ” means the domain of intentional application of the theory. The Sneedean school analyzed the theoretical and non-theoretical division, the division of typifications (or patterns) and laws, the problem of reduction and emergence of theories, and the unity or diversity of sciences with their approach of model theory and got successful results, while the Received View made no progress with any of those problems. It is very interesting that some philosophers want to develop this model-theoretic approach in philosophy of science and thus form the “new Vienna school”.

## 6 Conclusion

What is the relationship between the three kinds of model approaches introduced above? We believe that the state-space model approach expresses the behavioral trajectory of things in the state spaces and thus easily provides “icon models”,

whereas the set-theoretical model talks about the quantitative relation of things and is useful to provide mathematical models to ascertain natural laws, and the Sneedean model classes synthesize these two to provide icon models as well as mathematical models.

What does model mean? It is easy to be confused. According to the above analysis, icon models, mathematical models, state space models, or set-theory models may all be theoretical models, since they all fit with Tarski's definition that "a possible realization in which all valid sentences of a theory  $T$  are satisfied is called a model of  $T$ "[4].

What is model-based reasoning in science? Not only abductive reasoning and analogical reasoning are model-based reasoning, there is a general definition of model-based reasoning. Models are a isomorphisms or homomorphisms of the objective systems. There are two algebraic systems  $(X, \odot)$  and  $(Y, \oplus)$ . If  $g: x \rightarrow y$ , such that  $g(x_1 \odot x_2) = g(x_1) \oplus g(x_2)$ , that is, if it is possible from the laws of the source system  $(X, \odot)$  to infer laws in the objective system  $(Y, \oplus)$ , then that is model-based reasoning in science.

Therefore, a model is the carrier of knowledge, not just analogy and metaphor. Information in scientific experiments is typically expressed as a data model, and the blur data will becomes precise in modeling. Simulation models will replace step by step the experimental data to become the source of scientific research. Simulation and other dynamic models express the transformation of the states of systems to reflect nature. Theories and models are both composed of multi-dimensional mathematical spaces of states, mapping their functions to that of other systems, and thus may be studied through semantic concepts. In this sense, the model-theoretic approach is very promising in the philosophy of science, and is worthy of our attention.

## References

1. Suppe, F.: Understanding Scientific Theories: An Assessment of Developments, 1969-1998. *Philosophy of Science* 67, 102–115 (2000)
2. Hemple, C.G.: Formulation and Formalization of Scientific Theories. In: Suppe, F. (ed.) *The Structure of Scientific Theories*. University of Illinois Press (1979)
3. Hanfling, O.: Logical Positivism. In: *Routledge History of Philosophy*, p. 193f. Routledge (2003), Read more, <http://www.answers.com/topic/logical-positivism#ixzz1V9RplnkG>
4. Tarski, A.: Contributions to the theory of models. *Indagationes Mathematicae* 16, 572–588 (1954); 17, 56-64 (1955)
5. van Fraassen: *The Scientific Image*, p. 67. Clarendon Press, Oxford (1980)
6. Suppe, F.: *The Structure of scientific theories*. Edited with a critical introduction by Frederick Suppe, p. 221. University of Illinois Press (1977)
7. van Fraassen: On the Extension of Beth's Semantics of Physical Theories. *Philosophy of Science* 37(3), 328 (1970)
8. Monton, B.: Constructive Empiricism. *Stanford Encyclopedia of philosophy* (2008)



9. Suppe, F.: Theories, scientific. In: Craig, E. (ed.) *Routledge Encyclopedia of Philosophy*, Routledge, London (retrieved February 20, 2004)
10. Suppes, P.: *Introduction to Logic*, p. 249. van Nostrand Reinhold, New York (1957)
11. Torretti, R.: *The Philosophy of physics*, p. 415. Cambridge University press (1999)

# Cognitive Chance Discovery: From Abduction to Affordance

Akinori Abe

**Abstract.** In this paper, first I review the basic theories— concept and computational realization of abduction. Then I briefly review chance discovery which focuses on rare and novel events. In addition I briefly review the concept of affordance proposed by Gibson. By using the above concepts and techniques, a dementia care system inspired by affordance is proposed and discussed. Finally I introduce chance discovery based curation proposed by me. The dementia care system is discussed from the aspect of communication and chance discovery based curation.

## 1 Introduction

Abduction is one of sophisticated and intellectual process of human behaviour. By abduction we can determine unknown or less known matters. In addition, (computational) abduction can be applied to several applications such as design and planning. I have analyzed chance discovery by abduction [3, 4] and proposed chance discovery based system by abduction (e.g. [7, 5] etc.). In addition, recently I have discussed curation in the context of chance discovery and necessity of introducing a concept of curation to chance discovery applications [8].

Due to the advanced and innovative medical treatment, we are able to live longer. It will be happy to live long, but the other problems are caused by such long lives. One of the most famous problems is increasing patients who are suffered from cancer. It will be able to be overcome by the advancement of medical treatment and is a problem for individuals. Furthermore serious problem for a person and even for his/her family and surroundings will be dementia. It is the progressive decline in cognitive function due to damage

---

Akinori Abe

Faculty of Letters, Chiba University, Japan

e-mail: [ave@ultimaVI.arc.net.my](mailto:ave@ultimaVI.arc.net.my)

or disease in the body beyond what might be expected from normal aging. Dementia persons cannot reasonably live their lives. It is said that the current medical treatment cannot cure dementia completely. Even in the near future, it will be negative to cure dementia. Dementia is caused by problems in a brain. Accordingly, it is more difficult to cure dementia than cancer. Currently, some methods to delay the progress of dementia are proposed. For instance, a therapy room or house will be one of the solution to take care of dementia person [37]. Actually, it is rather a support system for dementia person's everyday life.

In addition, several researches and experiments are conducted to analyze the feature of dementia. Bozeat and Hodges showed affordance might give a certain support to (semantic) dementia persons of understanding (meanings of) objects [13, 22]. Actually, it covers a limited situation, but it would be better to introduce a concept of affordance to a dementia care. Affordance has been discussed in Artificial Intelligence or philosophy as well as in cognitive science. For instance, Magnani discussed manipulation of affordances in the abduction framework [26]. Thus strategies for dementia care can be discussed and built in the framework of affordance theory. Affordance theory is a natural processing in actual environments. In addition, affordance can be dealt with abduction framework and since affordance is not explicitly displayed but hidden in the environments. Accordingly, chance discovery [31] can be one of the strategies to deal with a dementia care problems.

In this paper, based on the above discussion and as an application of chance discovery based curation, a dementia care under the concept of affordance, abduction, and chance discovery is discussed.

As an introduction Sections 2 and 3 review several types of abduction and chance discovery. Section 4 illustrates the concept of affordance, abduction, and chance discovery which are discovery reasoning or knowledge processing. Section 5 reviews the feature of dementia. Section 6 proposes a dementia care system based on the concept of affordance. It will be discussed in the context of chance discovery. In addition, Section 7 discusses the dementia care system from the viewpoint of curation and communication as chance discovery. Section 8 concludes this paper.

## 2 Abduction

### 2.1 *Incomplete Knowledge Reasoning — Abduction and Induction*

In this section, as an incomplete knowledge reasoning (reasoning dealing with incomplete knowledge), I briefly introduce logical reasoning system — induction, and abduction.

The followings are a reasoning mechanism sequence of deduction.

- (1) Every man dies,
- (2) Enoch was a man;
- (3) Hence, Enoch must have died.

That is, if we know (1) and (2), we can conclude Enoch must have died (3). On the other hand, if a certain knowledge ((1) or (2)) is missing, we cannot conclude that “Enoch must have died.” In such a case, incomplete knowledge reasoning which Peirce classified as abduction and induction will be conducted.

Peirce classified *abduction* from a philosophical point of view as the operation of adopting an explanatory hypothesis and characterized its form.

- (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.

Where ‘reason (hypothesis)’ can not be easily assumed from A and C. In addition, he characterized *induction* as the operation of dealing and then testing a hypothesis by experiments.

- (1) Suppose that I have been led to surmise that among our coloured population there is greater tendency toward female birth than among our whites.
- (2) I say, if that be so, the last census must show it.
- (3) I examine the last census report and find that, sure enough, there was a somewhat greater proportion of female births among coloured births than white births in that census year.

Thus Peirce characterized abduction and induction as follows [32]:

- Abduction is an operation for adopting an explanatory hypothesis, which is subject to certain conditions, and that in pure abduction, there can never be justification for accepting the hypothesis other than through interrogation.

*Inference for (novel) discovery*

- Induction is an operation for testing a hypothesis by experiment, and if it is true, an observation made under certain conditions ought to have certain results.

*Inference for classification and learning, which are (generalized) discovery*

Thus although abduction and induction are categorized to an incomplete knowledge reasoning and discover something “new,” those which abduction discovers are rather different from those which induction discovers. If we want to discover general tendencies or classification induction will be better. On the other hand, if we want to discover something rare or novel, abduction will be better.

## 2.2 *Abductive Discovery*

Abduction can be applied to applications for new discovery. Very typical application of abduction will be discoveries or solutions in affairs. For instance, the following is a scene from a detective novel “A Study In Scarlet” by Arthur Conan Doyle.

*“Dr. Watson, Mr. Sherlock Holmes,” said Stamford, introducing us. “How are you?” he (= Holmes) said cordially, gripping my hand with a strength for which I (= Dr. Watson) should hardly have given him credit. “You have been in Afghanistan, I perceive.”...*

Of course, for the sudden utterance from a stranger which was astonishingly correct, Dr. Watson asked that “How on earth did you know that?” in astonishment. In fact, during several minutes when Holmes shook hands with Dr. Watson, Holmes concluded (=abduced) Dr. Watson had been in Afghanistan. He did not have any previous information of Dr. Watson, but with several observations he had such a conclusion. He illustrated his abduction procedure as below;

*Nothing of the sort. I (= Holmes) knew you (= Dr. Watson) came from Afghanistan. From long habit the train of thoughts ran so swiftly through my mind, that I arrived at the conclusion without being conscious of intermediate steps. There were such steps, however. The train of reasoning ran, ‘Here is a gentleman of a medical type, but with the air of a military man. Clearly an army doctor, then. He has just come from the tropics, for his face is dark, and that is not the natural tint of his skin, for his wrists are fair. He has undergone hardship and sickness, as his haggard face says clearly. His left arm has been injured. He holds it in a stiff and unnatural manner. Where in the tropics could an English army doctor have seen much hardship and got his arm wounded? Clearly in Afghanistan.’ The whole train of thought did not occupy a second. I then remarked that you came from Afghanistan, and you were astonished.*

In the above scene, Sherlock Holmes determined Dr. Watson’s vocation from the observation from Dr. Watson. Then Holmes guessed Dr. Watson’s situation. The process of the guesswork was not based on a “chance” but a very formal and logical inference. Of course, this process can be explained by abduction. Half of the above procedure are deduction to obtain (infer) observations for abduction and can be logically described as follows:

- 1) Dr. Watson is an army doctor ← medical type & with the air of a military man.
- 2) Dr. Watson is not colored ← wrists are fair.
- 3) Dr. Watson has just come back from the tropics ← face is dark & not\_colored.
- 4) Dr. Watson has undergone hardship and sickness ← haggard face & left arm has been injured.

5) Afghanistan  $\leftarrow$  English army doctor have much hardship and sickness & tropics.

That is, we can conduct deduction as follows:

- Observations: *medical\_type*, *wrists(fair)*, *face(dark)*, *haggard\_face*, *injured*, *air\_of\_a\_military\_man*
- deduction phase
  - $medical\_type \vee air\_of\_a\_military\_man \models army\_doctor$ .
  - $wrists(colored) \models colored$ .
  - $wrists(fair) \models not\_colored$ .
  - $face(dark) \vee not\_colored \models tropics$ .
  - $haggard\_face \vee injured \models hardship\_and\_sickness$

Then the rest of the inference process was logically performed based on observations (abduction). That is, Holmes generated Afghanistan as a hypothesis to explain various observations from Dr. Watson. In addition he used knowledge such as world situation in those days. The above inference process can be logically described as follows.

- abduction phase
  - Observations  $O$ : *hardship\_and\_sickness*, *tropics*, *army\_doctor*, *Englishman*
  - Facts  $F$ : knowledge sets in Holmes's brain
  - $\{Afghanistan, Malaysia, Russia, Japan, \dots\} \in H$

Actually, Holmes knew another feature of Afghanistan that Afghanistan is a harder place to live in than other countries in the tropics etc. Accordingly he could conclude (abduce) that Dr. Watson had been in Afghanistan. Thus hypothesis ( $h$ ) which is for “*in Afghanistan*” will be generated (selected) from  $H$ . The above is an inference by Sherlock Holmes (human inference). A computational inference will be illustrated in the following sections.

### 2.3 Computational Abduction

Abduction in the Artificial Intelligence field is generally understood as reasoning from observation to explanations, and induction as the generation of general rules from specific data. Sometimes, both types of inferences are regarded as the same because they can be viewed as being an inverse of deduction. For computation, Popper mechanized abduction as an inverse of deduction [35], although he seemed to distinguish abduction from induction. Muggleton and Buntine have formalised induction as an inverted resolution [28]. Both formalizations are realized as an inverse of deduction. In this paper, I will not discuss a relationship between abduction and induction. It was discussed in [1]. I will focus on a discussion on abduction.

Thus, abduction is usually used to find the reason (set of hypotheses) in a logical way to explain an observation. For instance, the inference mechanism of Theorist [33] that explains an observation ( $O$ ) by a consistent and minimal hypotheses set ( $h$ ) selected from a set of hypotheses ( $H$ ) is shown as followings.

$$F \not\vdash O. \quad (O \text{ can not be explained by only } F.) \quad (1)$$

$$F \cup h \vdash O. \quad (O \text{ can be explained by } F \text{ and } h.) \quad (2)$$

$$F \cup h \not\vdash \square. \quad (F \text{ and } h \text{ is consistent.}) \quad (3)$$

Where  $F$  is a fact (background knowledge) and  $\square$  is an empty clause. A hypothesis set ( $h$ ) is selected from a hypothesis base ( $h \in H$ ).

Thus, “reason” is usually selected from the knowledge (hypotheses) base. For instance, when Theorist is used for an LSI circuit design,  $F$  includes knowledge about the devices’ function and their connections, and the knowledge of other rules. In addition,  $H$  includes candidate devices and their candidate connections. If the relation between input and output of the circuit is given as an observation  $O$ , Theorist computes the name of devices and their connections as hypotheses  $h$ . Therefore, usual abduction requires a perfect hypotheses base from which a consistent hypotheses set is selected to explain an observation. Here, “perfect hypotheses base” means the hypotheses base that contains all the necessary hypotheses.

Clause Management System (CMS) was proposed by Reiter and de Kleer [36] and it was a database management system. Its mechanism is illustrated as follows:

When  $\Sigma \not\models C$ , if propositional clause  $C$  (observation) is given, CMS returns a set of minimal clauses  $S$  to clause set  $\Sigma$  such that

$$\Sigma \models S \vee C. \quad (4)$$

$$\Sigma \not\models S. \quad (5)$$

A clause  $S$  is called a minimal support clause, and  $\neg S$  is a clause set that is missing from clause set  $\Sigma$  that can explain  $C$ . Therefore, although CMS was not proposed as abduction, since from the abductive point of view  $\neg S$  can be thought of as an abductive hypothesis, CMS can be used for abduction.

In addition, I proposed Abductive Analogical Reasoning (AAR) which combines CMS-like abduction and analogical mapping. Details are shown in [2].

### 3 Chance Discovery

Chance Discovery is a discovery of chance, rather than discovery by chance. Ohsawa defined chance (risk) as “a novel or rare event/situation that can be conceived as either an opportunity or a risk in the future [31]”. It is naturally

understood that a chance, which is either known or unknown, includes possibilities to cause unfamiliar observations. It can also be said that a chance is an alarm like an inflation of money supply or a big difference between future (estimated, reserved) and current stock prices that will change the middle or long term economic situation (Japan, in 1990). We sometimes ignore such critical factors, because we cannot understand that they are important factors. This is because the results or the factors are exceptions, and rare or novel events.

Chance discovery is also characterized as an explanatory reasoning, however since “chance” is defined as unknown hypotheses, some techniques to deal with an empty or an imperfect hypotheses base are required. If so, such an inference mechanism as usual abduction (hypothetical reasoning etc.) is not sufficient to achieve chance discovery. Chance discovery needs an explanatory reasoning that can deal with an empty or imperfect hypotheses base.

In 2007 Taleb published “Black Swan” [39]. In the book, Taleb introduced a concept “Black Swan<sup>1</sup>” as an event with the following three attributes.

1. It is outlier, as it lies outside of the realm of regular expectations, because nothing in the past can convincingly point to its possibility.
2. It carries an extreme impact.
3. In spite of its outlier status, human nature makes us concoct explanations for its occurrence after the fact, making it explainable and predictable.

Thus Taleb discussed the similar event as a chance as black swan. I will not discuss black swan in this paper. The discussion was performed in [9].

## 4 Affordance

### 4.1 Affordance

Gibson ecologically introduced the concept of affordance for perceptual phenomena [19, 20]. It emphasizes the environmental information available in extended spatial and temporal pattern in optic arrays, for guiding the behaviors of animals, and for specifying ecological events. Thus he defined the affordance of something as “a specific combination of the properties of its substance and its surfaces taken with reference to an animal.” For instance, the affordance of climbing a stair step in a bipedal fashion has been described in terms of the height of a stair riser taken with reference to a person’s leg length [40]. That is, if a stair riser is less than 88% of a person’s leg length, then that means that the person can climb that stair. On the other hand, if a stair riser is greater than 88% of the person’s leg length, then that means that the person cannot climb that stair, at least not in a bipedal fashion.

---

<sup>1</sup> Black swans are native to Australia, but had never been seen in Europe.



For that Jones pointed out that “it should be noted also that this is true regardless of whether the person is aware of the relation between his or her leg length and the stair riser’s height, which suggests further that the meaning is not internally constructed and stored but rather is inherent in the person’s environment system” [23].

In the context of human-machine interaction Norman extended the concept of affordance from Gibson’s definition. He pointed out that “...the term affordance refers to the perceived and actual properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used. [...] Affordances provide strong clues to the operations of things. Plates are for pushing. Knobs are for turning. Slots are for inserting things into. Balls are for throwing or bouncing. When affordances are taken advantage of, the user knows what to do just by looking: no picture, label, or instruction needed” [30]. Thus Norman defined affordance as something of both actual and perceivable properties. Accordingly his interpretation has effectively been introduced to interaction designs.

Zhang categorized several types of affordance into the following categories [43]:

- Biological Affordance  
For instance, a healthy mushroom affords nutrition, while a toxic mushroom affords dying.
- Physical Affordance  
For instance, the flat horizontal panel on a door can only be pushed. Many of this type of affordances can be found in Norman [30].
- Perceptual Affordance  
In this category, affordances are mainly provided by spatial mappings. For instance, if the switches of the stovetop burners have the same spatial layout as the burners themselves, the switches provide affordances for controlling the burners. Examples of this type include the pictorial signs for ladies’ and men’s restrooms.
- Cognitive Affordance  
Affordances of this type are provided by cultural conventions. For instance, for traffic lights, red means “stop,” yellow means “prepare to stop,” and green means “go.”
- Mixed Affordance  
For instance, a mailbox, which is one of the examples used by Gibson, does not provide the affordance of mailing letters at all for a person who has no knowledge about postal systems. In this case, internal knowledge is involved in constructing the affordance in a great degree.

Thus since Gibson’s introduction, affordance has been widely discussed, and the other perspective and extensions have been added. Especially, it has been effectively introduced to interface designs after several extensions.

## 4.2 *Affordance, Abduction and Chance Discovery*

It is important to deal with rare or novel phenomena which might lead us to risk or opportunity. We call this type of activity as chance discovery and discuss theories and methods to discover such chances. A chance is defined as “*a novel or rare event/situation that can be conceived either as an opportunity or a risk in the future*” [31]. Thus it is rather difficult to discover a chance by usual statistical strategies. We adopt abduction and analogy (Abductive Analogical Reasoning [2] which can also be regarded as an extension of CMS [36]) to perform chance discovery [3, 4]. Where chance discovery is regarded as an explanatory reasoning for the unknown or unfamiliar observations, and a chance is therefore defined as followings:

1. **Chance** is a set of unknown hypotheses. Therefore, explanation of an observation is not influenced by it. Accordingly, a possible observation that should be explained cannot be explained. In this case, a hypotheses base or a knowledge base lacks necessary hypotheses. Therefore, it is necessary to generate missing hypotheses. Missing hypotheses are characterized as chance.
2. **Chance** itself is a set of known facts, but it is unknown how to use them to explain an observation. That is, a certain set of rules is missing. Accordingly, an observation cannot be explained by the facts. Since rules are usually generated by inductive ways, rules that are different from the trend cannot be generated. In this case, rules are generated by abductive methods, so trends are not considered. Abductively generated rules are characterized as chance.

Magnani also discussed application of abduction to chance discovery. Especially, he pointed out “manipulative abduction happens when we are thinking through doing and not only, in a pragmatic sense, about doing. So the idea of manipulative abduction goes beyond the well-known role of experiments as capable of forming new scientific laws by means of the results (the nature’s answers to the investigator’s question) they present, or of merely playing a predictive role (in confirmation and in falsification). Manipulative abduction refers to an extra-theoretical behavior that aims at creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices. The existence of this kind of extra-theoretical cognitive behavior is also testified by the many everyday situations in which humans are perfectly able to perform very efficacious (and habitual) tasks without the immediate possibility of realizing their conceptual explanation” [25]. Then he pointed out that “in dealing with the exploitation of cognitive resources and chances embedded in the environment, the notion of affordance, originally proposed by Gibson to illustrate the hybrid character of visual perception, can be extremely relevant. [...] In order to solve various controversies on the concept of affordance, we will take advantage of some useful insights that come from the study on abduction.

Abduction may fruitfully describe all those human and animal hypothetical inferences that are operated through actions which consist in smart manipulations to both detect new affordances and to create manufactured external objects that offer new affordances” [26]. Thus he suggests the application of abduction to detect affordances which can be regarded as chances embedded in the environment.

## 5 Dementia

Dementia is the progressive decline in cognitive function, such as memory, attention, language, and problem solving, due to damage or disease in the body beyond what might be expected from normal aging. In the later stages, dementia persons will not be able to recognize time (day of the week, day of the month, and year etc.), place, and person. Phenomena due to aging and dementia are different. For instance, for memory, aged person does not forget all of his/her experiences, on the other hand, dementia person forgets whole of his/her experiences. Dementia is roughly categorized to cortical and sub-cortical. For instance, several types of cortical dementia are reported such as Alzheimer’s disease. Except for the treatable types, there is no cure to dementia, although scientists are progressing in making a type of medication that will slow down the process. For instance, For the medication of Alzheimer, actions such as cheerful communication and proper stimulation are recommend [24]. For instance, some studies have found that music therapy which stimulates emotion as well as brain may be useful in helping patients with dementia [10]. Alternative therapies are also discussed for the care of Alzheimer’s disease and dementia [14, 15].

Bozeat and Hodges analyzed the feature of mapping between objects and their meaning for semantic dementia person from four factors — affordance, presence of recipient, familiarity, and problem solving [13, 22]. They showed very interesting results For instance, they pointed out “as a group, the patients did not achieve better performance on a subset of affordable objects when use of these was compared with a familiarity-matched subset of objects lacking such affordances. This absence of a general group benefit applied both to overall use and to the specific component of use afforded by the object’s structure.[...]it became clear that there was a reliable benefit of affordance on the specific components of use, but only for the most impaired patients.” They also pointed out “The impact of recipient, like affordance, was found to be modulated by the degree of semantic impairment. The patients with a moderate level of conceptual impairment demonstrated significantly better use with the recipient present, whereas the patients with mild and severe impairment showed no effect. [...] It was not surprising, therefore, to find that familiarity also influenced performance on object use assessments.”

These observations and analyses show that proper affordance might give a certain support to dementia persons understanding (meanings of) objects.

## 6 Dementia Care Inspired by Affordance

It is not possible to prepare all necessary things in every places. Sometimes an alternative or an extended usage of things will be necessary. For a proper and an extended usage of a thing, it is necessary to present proper information of it. At least, it is necessary to suggest such information. Sometimes it can be presented as a memorandum or a sign. In the other case, it can be received as hidden information inside of the thing. Actually it is not always necessary to provide such hidden information. For a progressive and promising system, it is not realistic to prepare all the necessary information to things or events. Sometimes such information is not always correct and may change in the future. For instance, it is ridiculous to attach a sign such as “You can sit here.” to tree stumps. Instead it is rather realistic to suggest information about its hidden functions.

In this section we discuss how to present such hidden information in dementia care situation. Such hidden information can be presented as certain stimuli in such situations. Because, as shown in the previous section, even for dementia person, if he/she receives certain stimuli, he/she sometimes achieve better performance. The problem is that what type of stimulus will be better to present and how to make it recognize. Actually such stimulus should be “afforded (selected from an environment)” by the user. That is, it can be regarded as an “affordance” in an environment. Accordingly we introduce concept of affordance to a dementia care system. Proper affordance might give a certain support to dementia persons understanding (meanings of) objects. Thus affordance is a fruitful concept for recognizing objects and using them as tools. According to Gibson’s definition, affordance is hidden in the nature and it should be accepted by us naturally. For instance, if an object’s upper side is flat and it has a certain height, the observer will be able to afford it as something to sit, rest or sleep. Of course, the level of affordance will be change according to observer’s acceptance ability. For a certain person a tree stump will function as a chair, but for the other person it will not. If they are able to regard a tree stump as a chair, it will be necessary to provide a proper guidance to discover an affordance as a something to sit.

For normal persons, it is not so difficult to provide such guidances. They can also understand analogy, so that they can extend the meaning to the other materials. For instance, after finding that a tree stump functions as a chair, they can also understand a wooden board or box can also function as a chair. That is, they can extend or map the meaning to the other situations. However, for dementia persons, it is not easy to provide a proper guidance with which they can afford the function of an object. Actually, for person who does not have common knowledge or context, it is also not easy to provide a proper guidance for affordance discovery. For them affordance is something rare or novel. Accordingly, it is rather difficult to be aware of “affordance” as an afforded matter. In therapy houses, there should be many things which are not able to properly used by dementia persons. In the case, it is necessary to

provide certain guidances to lead the user to the correct direction to use things properly. The simplest method will be to attach the name and usage of things. It will functions well for normal persons. However, for impaired persons, sometimes even such attachment will not function well. For them, it will be necessary to apply the other strategy to suggest or instruct the meaning or usage of things. For semantic dementia persons, it is observed that they did not achieve better performance on a subset of affordable objects when use of these was compared with a familiarity-matched subset of objects lacking such affordances. Therefore, when we design an environment for dementia persons, it is necessary to consider such unhappy situations. It is necessary to prepare specialized affordances to dementia person. Even if they can detect affordance, they might not understand what it will emerge.

For affordance, according to the Gibson's definition, an *Object* is observed and affordance is detected in the environment to understand its meaning. Then, when meaning is fixed, by using abduction framework, the affordance determination situation will logically be described as follows:

$$F \cup \text{affordance} \models \text{Object} \quad (6)$$

$$F \cup \text{affordance} \not\models \square \quad (7)$$

The above is described based on the formalization of Theorist [33].  $F$  is so called facts which involves fundamental knowledge in the world. The obtained affordance is consistent with  $F$  (equation (7)) and gives life (meaning) to the *Object*. Thus *Object* involves invisible *meaning* and by adopting discovered affordance, potential meaning appears. Therefore, in the above formalization, *meaning* does not appear explicitly.

However, in the above application, we would like to give a certain meaning to the *Object* explicitly. Though meaning exists inside of the *Object*, in this framework meaning is explicitly described. That is, meaning should be observed and affordance functions as a type of link to *Objects*. When meaning is fixed, the affordance determination situation will be logically described as follows:

$$\text{Object} \cup \text{affordance} \models \text{meaning} \quad (8)$$

$$\text{Object} \cup \text{affordance} \not\models \square \quad (9)$$

That is, affordance can be regarded as a hypothesis. We can select consistent affordance (equation (9)) in the environment (hypothesis base) to explain meaning. In addition, for understanding subset of or similar afforded objects (*Object'*), the affordance determination situation will be logically described as follows:

$$\text{Object} \cup \text{Object}' \cup M \cup \text{affordance} \models \text{meaning} \quad (10)$$

In fact, the above description is based on Goebel’s formalization of analogy [21].  $M$  is a mapping function from *Object* to *Object'*. That is, to understand the same meaning of the subset of or similar afforded objects, an additional mapping function  $M$  is required. Thus if  $M$  can be determined and the usage of *Object* is known, *Object'* can also be understood. In fact, for normal persons,  $M$  is easy to understand. However, for dementia persons, it is pointed out that it is rather difficult to understand and determine  $M$ . Then the issue becomes how to suggest a mapping function  $M$  as an additional hypothesis. For typical analogical mapping, objects in the source domain and the target domain are quite different. In fact, the typical analogical mapping is determined based on conceptual structure as pointed out by Gentner [16, 17, 18]. For instance, if we know about the water flow system where water flows from a place with greater pressure to a place with less pressure, we can guess or find the heat flow system where heat flows from a place with greater temperature to a place with less temperature. However, for the applications shown in this paper, a mapping function will not be so complex as typical analogical mapping. For the proposed application, expected situations are very simple. For instance, to give a hint (mapping function) of sitting on a wooden box to dementia person who could use a tree stump as a chair. In fact, the situation is generally structured, but for an application, we can only focus on an aspect such that the upper side is flat. This type of mapping will be one dimensional mapping and not so confusing. Thus theoretically a mapping function becomes simple. The above logical descriptions can be illustrated in Fig. 1.

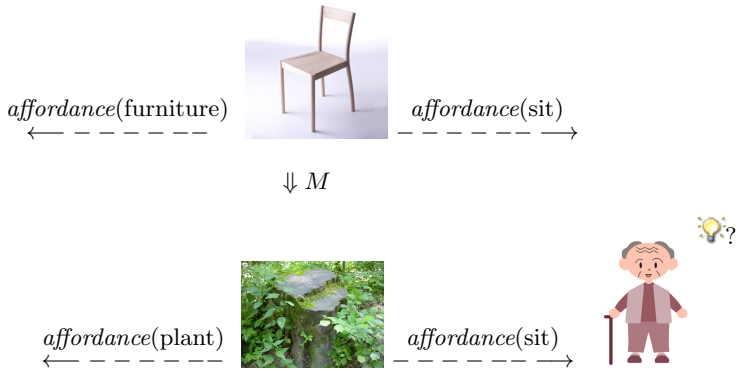


Fig. 1 Affordance: communication between human and environment

At last, the most important issue is how to suggest hidden information as affordance. An answer will be given in the following sections.

## 7 Curation and Communication as Chance Discovery

### 7.1 *Curation as Chance Discovery*

In chance discovery, interactions between human and computer to aid the discovery process is one of significant factors. Such interactions could be achieved by providing a user-friendly interface. For instance, visualization systems for making users aware of unconscious preferences [12, 27], an analogy game which varies a construction of concepts according to perceptions, categorizations, and areas of focus derived from the expertise of the observer [29], a deposit overflow determination system to prevent various financial crises [41], ISOR-2, a combination of case-based reasoning and statistical modeling system, which can deal with medical exceptions [42], and a web-based interactive interface which can check hidden or rare but very important relationships in medical diagnostic data sets [7] have been proposed in [6]. On the contrary, strategies of how to display chances have not been explicitly discussed in many applications. For a chance display strategy, I proposed a concept of curation for chance discovery [8]. A curation is a type of job in (art) museums and galleries, where curators remain up-to-date in the scholarly developments within their field(s), conduct original research and develop new scholarship that contributes to the advancement of the body of knowledge within their field(s) and within the museum profession as a whole [11]. In general, curators do not exercise communication with audiences except in special events such as gallery talks. In [8], I proposed the new definition of curation in chance discovery which is:

- Curation is a task to offer users opportunities to discover chances.
- Curation should be conducted with considering implicit and potential possibilities.
- Chances should not be explicitly displayed to users.
- However, such chances should rather easily be discovered and arranged according to the user's interests and situations.
- There should be a certain freedom for user to arrange chances.

Reviewing various applications in chance discovery, curation can also be regarded as an interactive communication between curators and audiences. Thus, it will be necessary to discuss “curation” from the aspect of “communication.”

Advertising communication also aims at offering certain information to audiences. Sometimes such information is implicitly expressed in an advertisement. As Pop summarized [34], advertising communication relies considerably on inferences and assumptions which help in proceeding towards eventual interpretations. Based on Grice's seminal theory of cooperative communication (cooperative principles, CP) and inferencing through a maxim of “filling in” or/and flouting, different interpretations could be accommodated by the linguistic theory. Pop extended her discussion by introducing

Relevance Theory (RT) [38], which explains hidden and additional information in advertisement.

Both curation and advertisement intend to deliver a certain concept to audiences. As easily guessed “communication” plays a significant role in curation and advertisement.

Affordance can be regarded as communication between human and environment, In the next section, a dementia care system inspired by affordance is discussed from the aspect of communication and curation.

## 7.2 Information Offering Strategies for Dementia Persons

In the Section 6, based on abduction, I reviewed the formalized concept of affordance based support system for dementia persons. In the formalizaion the most important relationship between an object and meaning is the last equation shown in Section 6. I review the equation below:

$$Object \cup Object' \cup M \cup affordance \models meaning \quad (11)$$

$M$  is a mapping function [21] from  $Object$  to  $Object'$ . That is, to understand the same meaning of the subset of or similar afforded objects, an additional mapping function  $M$  is required. Thus if  $M$  can be determined and the usage of  $Object$  is known,  $Object'$  can also be understood. In fact, for normal persons,  $M$  is easy to understand. However, for dementia persons, it is pointed out that it is rather difficult to understand and determine  $M$ . From the viewpoint of communication, if someone cannot understand or obtain the meaning of an object, it means that a communication link is missing between the object and the person and he/she cannot obtain any proper affordance given in the environment. In that case he/she needs certain hints to be aware of such affordance.

Thus the final issue is how to suggest hidden information as affordance. This type of information is usually hidden in the environment. Thus the proposed type of application can be discussed under the context of chance discovery. As I mentioned, chance discovery can be performed by a combination of abduction and analogy. Also as Magnani pointed out, affordance can be performed by a certain type of abduction. In the above, the concept of affordance is also described in the framework of Theorist that is hypothetical reasoning (limited version of abduction). Accordingly, all procedures can be described in abduction’s framework. In addition, it is happy for us that we can simplify our problems to one dimensional mapping. Of course, in this section, for the first step, a very simple case is discussed. For the actual usage, much more complex situation should be considered. My assumption is that such complex situation can be transformed to a combination of simple situations. To deal with complex situations, it is necessary to develop a mechanism to transform complex situation to a combination of simple situations



such as polynomial. Anyway, for such systems, chance discovery based curatorial strategies should be introduced to offer understandable mapping suggestion.

## 8 Conclusions

In this paper, first I reviewed abduction and chance discovery. They are basic techniques for applications discussed in this paper. Key techniques and concept in this paper are chance discovery, affordance and chance discovery based curation. In chance discovery we try to discover a novel or rare event/situation that can be conceived as either an opportunity or a risk in the future. The concept of affordance was ecologically introduced by Gibson for perceptual phenomena. It emphasizes the environmental information available in extended spatial and temporal pattern in optic arrays, for guiding the behaviors of animals, and for specifying ecological events. Currently we focus on the part of communication between human and environment. Based on the concept of affordance, I proposed a dementia person support mechanism in which functions of things can be implicitly suggested to dementia persons. It is based on abduction framework and performed under the context of chance discovery to determine affordance.

For the affordance determination, I adopt a concept of chance discovery based curation. Where chance display strategies are discussed. By a proper curation, it becomes even for dementia persons to determine better affordance.

Actually, I show a dementia care system but discussions in this paper can be applied to several applications such as a decision making support system.

## References

1. Abe, A.: On The Relation between Abductive Hypotheses and Inductive Hypotheses. In: Flach, P.A., Kakas, A.C. (eds.) *Abduction and Induction*, pp. 169–180. Kluwer (2000)
2. Abe, A.: Abductive Analogical Reasoning. *Systems and Computers in Japan* 31(1), 11–19 (2000)
3. Abe, A.: The Role of Abduction in Chance Discovery. *New Generation Computing* 21(1), 61–71 (2003)
4. Abe, A.: Abduction and Analogy in Chance Discovery. In: [31], ch. 16, pp. 231–248 (2003)
5. Abe, A., Ozaku, H.I., Kuwahara, N., Kogure, K.: Scenario Violation in Nursing Activities — Nursing Risk Management from the viewpoint of Chance Discovery. *Soft Computing Journal* 11(8), 799–809 (2007)
6. Abe, A.: Special issue on Chance Discovery. *International Journal of Advanced Intelligence Paradigms* 2(2/3) (2010)
7. Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: An Interface for Medical Diagnosis Support —from the viewpoint of Chance Discovery. *International Journal of Advanced Intelligence Paradigms* 2(2/3), 283–302 (2010)

8. Abe, A.: Curation in Chance Discovery. In: Proc. ICDM 2010 5th International Workshop on Chance Discovery, pp. 793–799 (2010)
9. Abe, A.: Relation between Chance Discovery and Black Swan Awareness. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS(LNAI), vol. 6882, pp. 495–504. Springer, Heidelberg (2011)
10. Aldridge, D.: Music Therapy in Dementia Care. Jessica Kingsley Publishers (2000)
11. American Association of Museums Curators Committee: A code of ethics for curators (2009), [http://www.curcom.org/\\_pdf/code\\_ethics2009.pdf](http://www.curcom.org/_pdf/code_ethics2009.pdf)
12. Amitani, S., Edmonds, E.: A Method for Visualising Possible Contexts. In: [6], pp. 110–124 (2010)
13. Bozeat, S., Ralph, M.A.L., Patterson, K., Hodges, J.R.: When objects lose their meaning: What happens to their use? Cognitive, Affective, & Behavioral Neurosciences 2(3), 236–251 (2002)
14. Cafalu, C.A.: The Role of Alternative Therapies in the Management of Alzheimer’s Disease and Dementia, Part I. Annals of Long-Term Care 13(7), 34–41 (2005)
15. Cafalu, C.A.: The Role of Alternative Therapies in the Management of Alzheimer’s Disease and Dementia, Part II. Annals of Long-Term Care 13(8), 33–39 (2005)
16. Gentner, D.: Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science 7, 155–170 (1983)
17. Gentner, D.: Analogical Inference and Analogical Access. Analogica, pp. 63–88. Pitman (1988)
18. Gentner, D.: The mechanisms of analogical learning. In: Similarity and Analogical Reasoning, pp. 199–241. Cambridge University Press (1989)
19. Gibson, J.J.: The Theory of Affordances. In: Shaw, R., Bransford, J. (eds.) Perceiving, Acting, and Knowing (1977)
20. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin (1979)
21. Goebel, R.: A sketch of analogy as reasoning with equality hypotheses. In: Proc. of Int’l Workshop Analogical and Inductive Inference. LNAI, vol. 397, pp. 243–253 (1989)
22. Hodges, J.R., et al.: The role of conceptual knowledge in object use evidence from semantic dementia. Brain 123, 1913–1925 (2000)
23. Jones, K.S.: What Is an Affordance? Ecological Psychology 15(2), 107–114 (2003)
24. Kasama A.: Dementia, <http://www.inetmie.or.jp/~Ekasamie/dementia.html>
25. Magnani, L.: Epistemic Mediators and Chance Morphodynamics. In: Abe, A., Ohsawa, Y. (eds.) Readings in Chance Discovery. International Series on Natural and Artificial Intelligence, Advanced Knowledge Intelligence, ch. 13, vol. 3, 140–155 (2005)
26. Magnani, L.: Chances, Affordances, and Cognitive Niche Construction: The Plasticity of Environmental Situatedness. International Journal on Advanced Intelligence Paradigms (2009) (to appear)
27. Maeno, Y., Ohsawa, Y.: Reflective visualization and verbalization of unconscious preference. In: [6], pp. 125–139 (2010)
28. Muggleton, S., Buntine, W.: Machine invention of first-order predicates by inverting resolution. In: Proc. of the 5th. International Workshop on Machine Learning, pp. 339–352 (1988)

29. Nakamura, J., Ohsawa, Y., Nishio, H.: An analogy game: toward cognitive upheaval through reflection-in-action. In: [6], pp. 220–234 (2010)
30. Norman, D.: *The Design of Everyday Things*. Addison Wesley (1988)
31. Ohsawa, Y., McBurney, P. (eds.): *Chance Discovery*. Springer (2003)
32. Peirce, C.S.: *Abduction and Induction*. In: *Philosophical Writings of Peirce*, ch. 11, pp. 150–156. Dover (1955)
33. Poole, D., Goebel, R., Aleliunas, R.: Theorist: A Logical Reasoning System for Defaults and Diagnosis. In: Cercone, N.J., McCalla, G. (eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*, pp. 331–352. Springer (1987)
34. Pop, A.: *Covert Communication in Advertising: A Case Study*, [http://www.upm.ro/facultati\\_departamente/stiinte\\_litere/conferinte/situl\\_integrare\\_europeana/Lucrari2/AnisoaraPop.pdf](http://www.upm.ro/facultati_departamente/stiinte_litere/conferinte/situl_integrare_europeana/Lucrari2/AnisoaraPop.pdf)
35. Pople Jr., H.E.: On The Mechanization of Abductive Logic. In: *Proc. of IJCAI 1973*, pp. 147–152 (1973)
36. Reiter, R., de Kleer, J.: Foundation of assumption-based truth maintenance systems: preliminary report. In: *Proc. of AAAI 1987*, pp. 183–188 (1987)
37. Sloane, P.D., et al.: The Therapeutic Environment Screening Survey for Nursing Homes (TESS-NH): An Observational Instrument for Assessing the Physical Environment of Institutional Settings for Persons With Dementia. *Journal of Gerontology: Social Sciences* 57B(2), S69–S78 (2002)
38. Sperber, D., Wilson, D.: *Relevance*, 2nd edn. Blackwell (1995)
39. Taleb, N.N.: *The Black Swan*. Allen Lane (2007)
40. Warren, W.H.: Perceiving affordances: Visual guidance of stair-climbing. *Journal of Experimental Psychology: Human Perception and Performance* 10, 683–703 (1984)
41. Yada, K., Washio, T., Ukai, Y.: Modeling Deposit Outflow in Financial Crises: Application to Branch Management and Customer Relationship Management. In: [6], pp. 254–270 (2010)
42. Vorobieva, O., Schmidt, R.: Case-Based Reasoning to Explain Medical Model Exceptions. In: [6], pp. 271–282 (2010)
43. Zhang, J., Patel, V.L.: Distributed cognition, representation, and affordance. *Cognition & Pragmatics* (2006)

# A Proposal on Belief, Abduction and Interpretation

Claudio Pizzi

**Abstract.** The paper starts from the claim that every assertion of belief may be analyzed into an assertion about a counterfactual surprise of the believer in front of a dissonant knowledge. The notion of a counterfactual surprise can be usefully related to the well studied notion of Shackle's potential surprise. In §2 the author stresses the difference between explanation and abduction on one side and interpretation on the other, maintaining that interpretation expresses what the interpreter believes to be the best explanation of the interpreted fact. The analysis which is proposed takes for granted the classical distinction between *doxa* and *episteme*, i.e. between a subjective and an objective dimension of the epistemic enterprise. However, in §1 belief is defined by making reference to knowledge, so reversing a relation which has been established by a deeply rooted philosophical tradition.

§1. The aim of this note is developing some reflections inspired by a basic intuition about belief statements which may be synthesized as follows :

- 1) Belief statements have an essentially counterfactual character
- 2) They presuppose the notion of knowledge and are definable in terms of it
- 3) They have to do with the notion of surprise

To be clearer, the analysis of belief which we want to take as a starting point is provided by the following definition:

(Def *Be*)  $x$  believes that  $A \Rightarrow_{\text{Df}} x$  would be surprised if  $x$  were to know that not- $A$

---

Claudio Pizzi

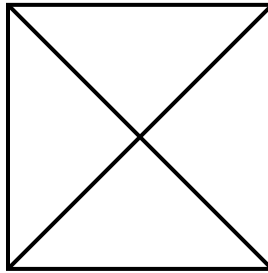
Dipartimento di Filosofia e Scienze Sociali, Università di Siena, Via Roma 47, 53100 Siena  
e-mail: pizzic@msn.com

A colloquial equivalent of the right part of Def *Be* which will be used in what follows is “ $x$  would be surprised to know that not- $A$ ”<sup>1</sup>. Of course from Def *Be* we derive the following equivalence: “ $x$  believes that not- $A$  if and only if  $x$  would be surprised to know that  $A$ ”. We will say that  $x$  *excludes that A* to mean the  $x$  believes that not- $A$

A square of oppositions for belief has then the shape indicated in Fig. 1:

**$x$  believes that  $A$**

**$x$  excludes that  $A$**



**$x$  does not exclude  
that  $A$**

**$x$  does not believe  
that  $A$**

**Fig. 1**

The two lower corners of the square deserve attention. According to the proposed definition, “ $x$  does not exclude that  $A$ ” means “ $x$  would not be surprised to know that  $A$ ”, and this is obviously implied by “ $x$  believes that  $A$ ”. How to render the left lower corner by using simply the word “belief”? By suitable substitutions we should use the phrase “ $x$  does *not* believe that *not*  $A$ ” (i.e. the dual of “ $x$  believes that  $A$ ); however, the meaning of this sentence containing two negations is cumbersome. As a matter of fact, “not believing that not” appears to be express a weak form of belief, but it is difficult to find a couple of words which in Italian or English grasp the exact distinction between the two senses of belief. In front of this terminological gap an approximation to what is meant is offered by the distinction between “believing that  $A$  is true” and “believing that  $A$  is possible”, where of course the former phrase implies the latter.

<sup>1</sup> A variant of this theory could be given by replacing knowledge with information. The paraphrase in this case then would be: “ $x$  would be surprised if  $x$  were to receive the information that  $A$ ”. The only inconvenience of this analysis is that the concept of information is more equivocal than the notion of knowledge, which may be represented by an axiomatizable modal operator.

The distinction between a weak and a strong notion of belief is not a novelty in philosophical literature. Sometimes a distinction is made between beliefs whose refutation destroys the whole epistemic systems to which they belong, and beliefs which have not such catastrophic consequences<sup>2</sup>. In our approach, however, the distinction relies on the psychological reaction which the subject is supposed to have in front of a new information: what is at stake here is, in fact, the difference between being surprised by something and *not* being surprised by something. Not being surprised, say, to know that a biased coin will give head is compatible with not being surprised that it will give tail, while, for every A, being surprised by A is incompatible with being surprised by not-A<sup>3</sup>.

A non-trivial difficulty of the present approach concerns the properties of the conditionals which are implicit in assertions of counterfactual surprise. Suppose in fact we use the symbol “>” for the conditional operator and the symbol *S* for “*x* is surprised”. Then the interrelation between the belief statements is described by the square in fig.2. The problem is that if “>” stands for the material conditional  $\supset$ , the strict conditional  $\rightarrow$  or the Stalnaker-Lewis conditional  $\square\rightarrow$ , there is no way to justify subalternance, i.e. the fact that  $C > S$  implies  $\neg(\neg C > S)$  and  $\neg C > S$  implies  $\neg(C > \neg S)$ .

However, there is a non-standard logic of implication, i.e. so-called *logic of consequential implication*, which grants that  $C > S$  implies both  $\neg(C > \neg S)$  and  $\neg(\neg C > S)$ . Such laws are called “Boethius’ Thesis” and “Secondary Boethius’ Thesis”<sup>4</sup>.

The semantic idea which is at the basis of the “consequentialist” view of conditionals is that the truth of a conditional depends on a consequential nexus between the clauses. In this approach the correct formal representation of negation is controversial. The surface form of “If tomorrow rains I will not be surprised” is represented by  $R > \neg S$ , but from a consequentialist viewpoint one cannot say that this conditional is true since we cannot say that the non-surprise  $\neg S$  is a consequence of raining via some logical or physical law. It is better to hold that in such cases a consequence relation is lacking between the clauses, so the sound formal representation is not  $R > \neg S$  but  $\neg(R > S)$ <sup>5</sup>.

<sup>2</sup> See e.g. Carnielli and Pizzi [2008], p. 201.

<sup>3</sup> We omit treating here the sense of belief in which the object of belief is not something which cannot be classified as true, false, possibly true or possibly false. Belief in fact may be concerned with normative or aesthetic evaluations. For instance I could say: “I believe that animals should be respected” or “I do not believe that *La Dolce Vita* is a masterpiece”. Here it seems to be improper speaking of “surprise to know”, since knowledge implies truth. Knowledge, however, could be here referred to some more complex object. The first statement implies that I would be surprised to know that there is general approval of some moral code implying cruelty to animals. In the second statement, where a weak sense of belief is involved, I mean that I would not be surprised to know that there is a disagreement on the commonplace view that *La Dolce Vita* is a masterpiece.

<sup>4</sup> See for instance Pizzi [2004].

<sup>5</sup> On the problem of formalizing negation in conditionals see Pizzi [1981].

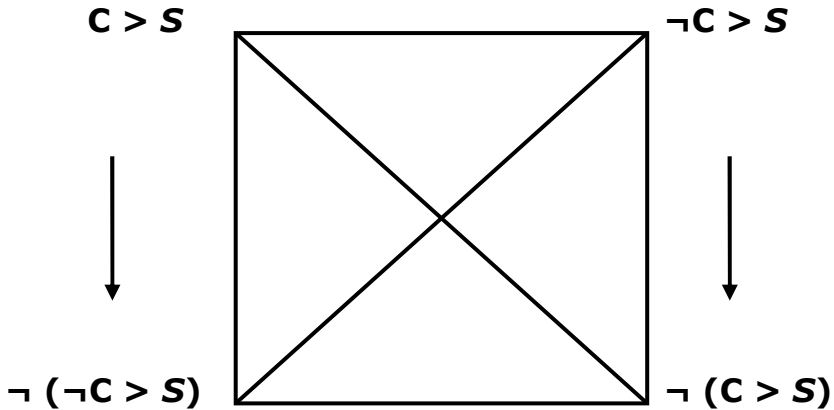


Fig. 2

§2. It is worth while noticing that the preceding analysis of belief preserves the famous Frege-Quine paradoxes of belief. For instance, from the true premises

a1) Smith believes that Rio de Janeiro is the capital of Brazil

a2) Brasilia = the capital of Brazil

one should draw the false conclusion

a3) Smith believes that Rio de Janeiro is Brasilia.

An analogous paradox is provided by applying to a1) the preceding analysis of belief in terms of counterfactual surprise. a1) in fact is paraphrased into

a\*1) Smith should be surprised to know that Rio de Janeiro is not the capital of Brazil

ad by a\*1) one should derive by the truth a2)

a\*3) Smith should be surprised to know that Rio de Janeiro is not Brasilia.

Now a\*3) is clearly false, since no surprise is caused to a normal subject by learning that cities with different names are different cities.

Also Moore's celebrated paradox "It rains, but I do not believe it" is preserved under the proposed analysis. According to it, in fact, Moore's sentence would amount to "it rains, but I would not be surprised to know that it does not rain": this is not a contradiction but surely a "logical oddity", as Moore's paradox is usually considered to be.

The preceding remarks give some supplementary plausibility to the definition of belief in terms of counterfactual surprise. However, a possible source of perplexity embodied in this proposal is provided by the fact that the relation between knowledge and belief which is presupposed in it subverts the traditional relation between these two notions. Beginning from Plato's *Theaetetus*, the classical position is defining knowledge in terms of belief, not *vice versa*. Attempts in this reverse direction have been sometimes proposed, but have not been successful. In a paper entitled *Belief as Relative Knowledge*, for instance, John Bacon (see [1975]) introduces the definition of belief in this way:

(Def Be)  $\text{Be}A =_{\text{Df}} K(\nabla \vee A)$

Where  $\text{Be}A$  stands for “ $x$  believes that  $A$ ”,  $K$  stands for “ $x$  knows” and  $\nabla$  is a propositional constant for “ $x$  is mistaken”, “I believe that  $A$ ” means “I know that  $A$  is true, unless I mistake”. The inconvenient of this original proposal is that if I believe that I am mistaken, this is coincident with the fact that I know it. In fact, by substituting  $\nabla$  to  $A$  we have the equivalence

$\text{Be}\nabla \equiv K(\nabla \vee \nabla) \equiv K\nabla$

On the other hand, if  $\Delta$  stands for  $\neg\nabla$  (so it has the meaning of “ $x$  is not mistaken”), knowledge can be defined in terms of belief in this way:

(Def K)  $\text{K}A =_{\text{Df}} \Delta \wedge \text{B}A$

(“ $x$  knows that  $A$ ” means that  $x$  believes that  $A$  and  $x$  is not mistaken).

The main philosophical difference between Bacon’s approach and the present approach is that “being mistaken” or “being not mistaken” is something which does not depend on the mind of the believer (it is, in other words, an objective fact), while “being surprised” or “being not surprised” is something which is subject-depending. So the present analysis does not question the traditional distinction between the objective dimension of knowledge (*episteme*) and the subjective dimension of belief (*doxa*), but relies on it in a well defined sense.

Three further remarks are in order:

1) If I know that  $A$ , I would be surprised to know that  $A$  is false. This implication is intuitively sound: If I know  $A$ ,  $\neg A$  is incompatible with what I know, so to know  $\neg A$  would be a source of surprise. So knowledge implies belief, even, if the correct reconstruction inside a formal calculus is surely a complicated matter.

2) The properties of the logic of belief which result from the proposed analysis turn out to be dependent on the presupposed properties of counterfactual conditionals. Such properties are established by the background axioms and by the language in which they are formulated. Neglecting details about axioms, it is not obvious, for instance, that the rules of formation of the background language admit iteration of conditional operators, even if the most common assumption is that nesting of conditional operators is to be allowed by formation rules<sup>6</sup>. In other words, we may meaningfully say such things as “I know that you believe”, “I believe that you know”, “I believe that you believe” and so on. For instance, “I believe that Johnny believes that  $A$ ” is translated into “I would be surprised to know that Johnny would not be surprised to know  $\neg A$ ”.

What could be said about the often discussed thesis

(BB): “If  $x$  believes  $A$ ,  $x$  believes that  $x$  believes  $A$ ”?

In our approach (BB) is translated into “If  $x$  were surprised to know that  $\neg A$ ,  $x$  would be surprised to know that  $x$  is not surprised to know  $\neg A$ ”. This appears to be intuitively true, even if its derivation in a logical system may need special assumptions about the concept of knowledge and the concept of surprise.

3) Last but not least, the proposed definition opens the possibility to express degrees of belief as degrees of counterfactual surprise. Such degrees may be

<sup>6</sup> For a different option see for instance Del Grande [1987]. For the importance of admitting iteration of conditional operators see Pizzi [1999] and [2007].



expressed by a metric scale or by a non-metric scale. To believe strongly (weakly) that A means to have a high (weak) degree of counterfactual surprise due to A. More specifically, to believe in A to extent  $x$  means to have a counterfactual surprise to extent  $x$  in knowing not-A.

§3. The last point mentioned in §2 suggests a connection between the analysis introduced in §1-2 and the ideas of scholars who introduced the concept of surprise as a key to the analysis of reasoning with uncertain conclusions. As a matter of fact, the notion of surprise has become a key-notion in epistemological reflections of the last decades. Surprise may be qualified as the emotional response which takes places when a subject receives information that does not cohere with his/her current representations. It has been identified as one of the six basic universal emotions (Ekman [1992]), and is associated with a distinct bodily reaction over widely divergent cultures. P. Thagard in [2006] highlights the particular adaptive value of surprise, stressing that surprise leads one into a cycle of questioning and, possibly, discovery.

An obvious reference in this connection is to the important contribution of the economist R. Shackle. Referring to the supposed degree of possibility of an event, Shackle says: «It is the degree of surprise to which we expose ourselves, when we examine an imagined happening as to its possibility, in general or in the prevailing circumstances, and assess the obstacles, tensions and difficulties which arise in our minds when we try to imagine it occurring, that provides the indicator of degree of possibility. This is the surprise we *should* feel, if the given thing *did* happen; it is *potential* surprise»<sup>7</sup>.

The preceding description of potential surprise makes it clear that Shackle has in mind something which is akin to what we define here as counterfactual surprise. In Shackle's theory, the belief in  $h$  is the degree of disbelief in  $\neg h$ : in symbols,  $b(h) = d(\neg h)$ . The degree of disbelief  $d(\neg h)$  expresses the potential surprise of not- $h$ .

If potential surprise is the same as disbelief, one could suppose that it has the same behaviour of improbability, so that, if Pr represents the standard (Kolmogorov) probability function,  $d(A)$  should be equal to  $\text{Pr}(\neg A)$  or  $1 - \text{Pr}(A)$ . This impression is however wrong. If  $d(A)$  is a rational number expressing the degree of potential surprise of A, the sum of  $d(A)$  and  $d(\neg A)$  may be far from 1. It may happen in fact that, for lack of information, both A and  $\neg A$  are both surprising: for instance in some strange case of death both the hypothesis of an accident and the hypothesis of a murder or suicide are surprising. The so-called principle of multiplication which works for probability is also implausible. For instance, if two witnesses  $a$  and  $b$  independently tell the same story, the degree of surprise that both lie in this special example is intuitively higher than the degree of surprise produced by the falsity of one of the two testimonies, say  $a$  :

$$(DS) d(\neg A \wedge \neg B) > d(\neg A)$$

But  $\text{Pr}(\neg A \wedge \neg B) \leq \text{Pr}(\neg B)$  holds by Kolmogorov axioms for every instance of A and B, so  $d$  and Pr clearly have divergent properties.

The principle which is more generally formulated as

$$(d\wedge) d(A \wedge B) \geq \min(d(A), d(B))$$

is found in several different theories of uncertain reasoning: the most known

---

<sup>7</sup> Shackle, [1961],p.68

are J. Cohen's neoBaconian inductive probability and G. Shafer's theory of evidence<sup>8</sup>.

A criticism which Shackle took in consideration the following. If I do the wrong number of telephone, this is not a fact which I would call surprising; but nonetheless I am convinced that I got the right number. So one could argue that the two notions of surprise and belief are independent notions. But Shackle had a reply: "I can attach zero potential surprise to getting a wrong number, but also zero potential surprise to getting the right one. Thus both my degree of belief in getting a wrong number, and my surprise if I do, will be zero. It is when we interpret "degree of belief" in some sense resembling subjective distributional probability that we can find no basis for, or meaning in, a formal reconciliation of the two concepts of belief and surprise" ([1961],p.72).

The translation of Shackle's counterargument into our conditional language is not a trivial problem. When Shackle speaks of "zero potential surprise" of  $W$  and  $R$  (where of course  $W = \neg R$  and  $R = \neg W$ ), this should be represented in his language by  $d(W) = 0$  and  $d(R) = 0$  respectively. In our non metric conditional language the two concepts should be rendered as  $W > \neg S$  and  $R > \neg S$ . However, for reasons which have already been exposed, in our language the best rendering of the two propositions is given by  $\neg(W > S)$  and  $\neg(R > S)$ , which both can belong to the opposition square described at p.174. Being subcontraries, the two statements cannot both be false but can both be true, exactly as Shackle say.

A final comment on this question is that conditional language turns out to be more analytical than standard logical language, even if endowed with a metric for potential surprise. The distinction between strong and weak belief, for instance, is not clearly treatable in terms of degree of potential surprise. On the other hand, even if nothing prevents extending the conditional language with metric operators so to allow statements of form, say,  $W > d(W)=0$ , it is not obvious that the operator  $d$  should apply not only to truthfunctional statements but to simple and iterated conditionals.

§ 4. It is of some interest here to remark that when Peirce defines the notion of abduction he uses the notion of surprise.

(PA) "The surprising fact,  $F$ , is observed; But if  $H$  were true,  $F$  would be a matter of course. hence, there is reason to suspect that  $H$  is true"<sup>9</sup>.

The notion of surprise used by Peirce should be understood and carefully studied in the context of his system of thought. It is clear anyway that in (PA) Peirce intends that a fact  $F$  is surprising when it is unexpected or - more plausibly-unexplained<sup>10</sup>. Being  $F$  unexplained, we look for an explanation of it, and the abductive process stops when some hypothesis  $H$  provides a natural explanation of  $F$ . As a matter of fact, Peirce seems to give to the word "surprising" a sense which

<sup>8</sup> See Cohen [1977] and Shafer [1976].

<sup>9</sup> Peirce, C.S. [ 1935 ], 12, 5.189 .

<sup>10</sup> According to the well-known Hempel Symmetry Thesis, prediction and explanation may be converted in all contexts. For our purposes it is enough to say that explanation implies prediction, so being unexpected (surprising) implies being unexplained.

refers to an objective lack of an explanation and not to the mental or psychological state of some specific subject.

The relation between explanation and surprise has been explored by various epistemologists, but with results which are open to the charge of subjectivism. According to P. Gärdenfors, for instance,

a) The role of the *explanans* is to convey information about the *explanandum*

b) The main effect of the *explanans* is that the degree of surprise of E is decreased. In other words, to explain something is to reduce its degree of surprise<sup>11</sup>.

Gärdenfors sees the progress of science and of logic itself in terms of dynamics of belief, so making the notion of belief the central epistemological concept<sup>12</sup>. This viewpoint belongs to a common trend in postpositivistic thought, where science is often seen as a set of beliefs which, as such, is not in principle different from any set of beliefs of non-scientific nature.

The line of thought followed here is different since it does not intend to question the classical distinction between *doxa* and *episteme* or, in other words, between what is a subjective opinion (belief) and what is objective knowledge. The only relation between such notions, according to the view held here, is that belief is seen as depending on some counterfactual state of knowledge. Explanation and abduction, as are here intended, do not depend on belief or surprise. We will also take for granted that explanation is essentially a relation between an *explanans* and an *explanandum* (in Hempel's classical sense), while abduction is inference to the explanation which is "the best" in some well defined sense. The definition of such concepts may be refined in various ways but, as they are intended here, it does not need any reference to the mind of any subject.

§5. The hints we have introduced in the preceding sections allow developing some remarks about the concept of interpretation, an important notion which unfortunately is used or abused in a plurality of meanings. The first treatment of interpretation may be found in Aristotle (*De Interpretatione*). Here Aristotle says that interpretation is the reference of signs to concepts (*affections of mind*) and of concepts to things. For Aristotle and his Medieval followers interpretation is a mental activity, and some contemporary theoreticians, such e.g. as Ogden e Richards in [1923], endorse such mentalism.

The name of Peirce, the founder of modern semiotics, is again to be recalled in connection with the notion of interpretation. According to him interpretation is a three-place relation which involves a sign, the object to which the sign refers and an interpretant (an interpretant being in its turn a sign which yields the relation between the first and the second term)<sup>13</sup>. In this antimentalistic perspective the mental act of interpreting is replaced by the habit of action, i.e. by the regular reply which the interpreter associates to the sign).

<sup>11</sup> Gärdenfors [1990], p.102. In his view, "the surprise value of E is inversely related to the degree of belief associated with E in P-E " (p.109), where P-E is the epistemic state P contracted by the elimination of E.

<sup>12</sup> Gärdenfors [1985]

<sup>13</sup> See Magnani [2009], p. 169.

The plurality of meanings in which the word “interpretation” is actually used is impressive. Interpreting a musical score means transforming certain signs belonging to written musical language into a set of sounds, where different human interpreters may perform such a transformation in different ways. The interpretation of a formal language is an assignment of meaning to the constants and variables of the language. In the same vein, one speaks of interpretation of Quantum Mechanics intending the meaning which different scientist may associate to quantum formalisms.

There are also senses in which the term “interpretation” is not applied to symbolic or linguistic facts. One speaks for instance of “interpretation of the dreams of Mr. Rossi”, “interpretation of Fascism”, “interpretation of French Revolution”. Here interpretation seems to be what appears to be “the best explanation” of a phenomenon, or at least the best explanation according to some selection of data operated inside some given set of informations. A paradigmatic case of interpretation in this sense is provided by the radar operator who interprets a sequence of “plots” appearing on his screen as the movement of an airplane.

What seems to be clear is that when some linguistic or extralinguistic object receives an interpretation, this implies that at least another interpretation is in principle possible. If, for instance, it happens that  $x$  translates the string of words (CN) *Cane nero* as “black dog” and  $y$  translates it as “Sing, Nero!”, we are inclined to say that both offer a different interpretation of CN and that each of them is not only a translator but an interpreter of CN.

In what follows we want to give priority to the notion of interpretation as something applied to facts, not to strings of signs. More exactly, we want to maintain that the privileged sense of interpretation is provided by the following analysis, where  $x$  is a human subject:

(I)  $x$  interprets  $Q$  as P

means

(I') according to  $x$ , the best explanation of  $Q$  is P

or equivalently

(I'')  $x$  believes that the best explanation of  $Q$  is provided by P.

This theoretical option opens the problem of reducing other senses of interpretation to what we hold to be the primary one. This is surely not an easy task and can be the goal of a complex program of inquiry. To begin with, here we can simply suggest that if “interpreting” implies “translating”, to say that  $x$  interprets  $Q$  as P means that  $x$  believes that the best translation of  $Q$  is provided by P. Suppose for instance that  $x$  knows only two languages, Italian and English: he will interpret the phrase “*Cane nero*” as “*black dog*” in the sense that it is the best translation he knows of this string of words. But if he knows only Latin and English, he will interpret “*Cane nero*” as “*Sing, Nero!*”. In case he knows Latin, Italian and English he will make a choice among the two alternative translations and will normally find that one of the two possible translations of “black dog” is the best one.

The basic problem, however, is how to pass from “translating from  $x$ 's viewpoint” to “explaining from  $x$ 's viewpoint”. This step can be done in several ways. As it is often said, explaining means giving a reply to a why-question. Suppose that I do not

know Latin nor latin languages and suppose that someone grants to me that beyond any reasonable doubt the couple of words “*Cane nero*” is a meaningful string of signs. I may then ask not “Which is the translation of “*Cane nero*”? but “Why is “*Cane nero*” meaningful<sup>14</sup> ?” Reply: “it is meaningful since there is a language in which it has a meaning, and in our language such meaning is given by the meaningful phrase “Black dog””. Here the *explanandum* follows by the *explanans* by an argument which obeys the so-called Nomological-Deductive schema, even if no natural law is actually used in the derivation. If furthermore my interlocutor believes that, compared with other possible explanations, the quoted explanation is also the *best* explanation, the interlocutor is offering to me an *interpretation*, and more exactly his *interpretation*, of the string of signs “*Cane nero*”.

The string “*Cane nero*” may be or not be the product of intentional activity. In most interesting cases, however, interpretation concerns the product of an intentional activity, as for instance when one speaks of the interpretation of a norm of the penal code. In this case the why-question to which we are asked to reply is: “Why did the law-maker introduce that norm?” or also: “Why did the law-maker use that formulation of the norm in place of some different formulation?” In such case what is believed to be the best explanation of the *explanandum* is provided by indicating what is believed to be the best hypothesis about the intentions of the law-maker. This amounts to what is usually called “interpretation of a norm”.

The proposed analysis allows understanding the important phenomenon called “*interpretation of sense-data*”. It is well-known that post-positivistic philosophy put emphasis on so-called “*theory laden data*”( i.e. data interpreted in the light of a theory) and that it made a commonplace of the thesis that the evidence which supports scientific theories consists in theory-laden data. Now an original feature of Peirce’s philosophy has been the view that there is an abductive element in perception, i.e. that perception is a complex activity involving inference about the explanatory causes of sense-data. Indeed, to quote again Peirce, “abductive inference shades into perceptual judgment without any sharp line of demarcation between them” ([1955], 304). A prominent position in this realm of abductive inferences is given to what is called “ visual abduction”<sup>15</sup>, but according to Peirce abduction is involved in all sensorial activity.

The role of interpretation of sense impressions becomes evident in Gestalt phenomena, the most famous of which are the so called-Gestalt effects<sup>16</sup>. In this connection it is useful to recall here a famous imaginary example suggested by N.R. Hanson in *Patterns of Discovery* ([1958]). Suppose Tycho and Kepler look at sunrise from a hill. In a sense surely they see the same thing (their eyes receive the same stimulus; this is what we will call *see-1*); but in another sense (the sense of *see-2*) they “see” different things. Tycho sees-2 the sun going up from the horizon, while Kepler sees-2 the horizon going down.

<sup>14</sup> It is often remarked that there is ambiguity in why-questions. It may mean “which is the cause of...?”, “what is the explanation of ...?” but also “how do you know that...?”. The latter meaning seems to be the most suitable to the example.

<sup>15</sup> The literature on this point is widespread. See for instance Magnani [2009], p. 268 ff.; Moriarty [1996] and the dissertation Fiorelli [2005].

<sup>16</sup> In Pizzi [2006] it is held that Gestalt phenomena and Gestalt effects occur not only in the field of perception but also in inferential activity.

According to Hanson and most post-positivists, every scientist, being conditioned by what Kuhn calls the background paradigm, cannot avoid seeing phenomena in the sense of see-2. More generally, according to this school of thought there is no perception which is not theory-laden. The controversy on this thesis engaged in the last decades of the last century the most important epistemologists of that time, and we need not recall it here. We may simply observe that it is normally possible to reduce any proposition  $A^T$  which is “theory laden” to some proposition  $A$  which, even if not completely neuter, does not show the dependence of  $A^T$  from some background theory  $T$ . This reduction could be called *derelativization*. It is understood that derelativization cannot be unlimited, in the sense that sooner or later in the process of derelativization one reaches a statement which meets the agreement of all people having a sound and normal sensorial apparatus.

In the case of the Tycho-Kepler example the derelativized  $A$  is given by a proposition stating that the distance between the sun and the horizon increases during the given interval of time. So the propositions that we formulated as Kepler’s and Tycho’s reports are actually interpretations of  $A$  in the before defined sense: each of the two astronomers states what is the explanation of  $A$  according to the theory which each of them believes to be the best theory, or in other words what is the best explanation of  $A$  from each one’s viewpoint.

So our claim is essentially that divergence of interpretation is a divergence of beliefs concerning what is the best explanation of a given fact, which is to say that divergence of beliefs concerns the result of an abductive inference. But here we can make a further step in the analysis by recalling the definition of belief we proposed at the beginning. According to what was proposed at p.173, Kepler’ belief, for instance, may be restated in the following way:

(SK) Kepler would have been surprised to know that the best explanation of his sense-impression was not given by the movement of the earth.

Tycho’s belief on the contrary might be restated as follows:

(ST) Tycho would have been surprised to know that the best explanation of his sense-impression was not given by the movement of the sun.

To give more plausibility to the preceding analysis we observe that the feeling of surprise is something which experimental psychologists record just in describing the reactions to Gestalt effects. If in the famous duck-rabbit experiment you initially “see” a rabbit, you will be surprised when someone informs you that the same drawing could be also seen as the representation of a duck; and *vice versa* if you initially “see” a duck. In other words, in both cases you will be surprised to know that what you think to be the best explanation of the sense-impressions you receive is not such, since there is another equally plausible explanation of your sense impressions<sup>17</sup>.

---

<sup>17</sup> One could find that the fact described in the sentence “ $A$  represents a duck” is not an uninterpreted fact or, in other words, that we should derelativize this fact (in the mentioned sense) in order to perform a correct abduction. But this operation can be performed in various way, for instance by reducing the given sentence to something as “the intention of the drawer was to draw a duck” or “there is set of rules of projections which associate a subset of the points of the drawing to a duck”.

The main point which results from the preceding observations is that interpretation belongs to the dimension of subjectivity, while explanation and abduction do not show this dependence. This does not exclude, of course, that interpretations and beliefs concerning some *datum* can be ordered according to some scale of rationality and reasonableness, or simply in function of the amount of agreement they receive by expert subjects. Furthermore, analysing interpretations in terms of surprise suggests that, since it is common to speak of degrees of surprise, one should give sense also to speaking of degrees at which the explanation of something (a text, a fact, a sense *datum*) is affected by the mediation of the interpreter himself.

## References

- Bacon, J.: Belief as relative knowledge. In: Anderson, A.R., Barcan Marcus, R., Martin, R.M. (eds.) *The Logical Enterprise*, pp. 189–210. Yale, U.P. (1975)
- Carnielli, W., Pizzi, C.: *Modalities and Multimodalities*. Springer, Berlin (2008)
- Cohen, J.: *The Probable and the Provable*. Oxford University Press (1977)
- Del Grande, J.P.: A first order conditional logic for prototypical properties. *Artificial Intelligence* 33, 105–130 (1987)
- Ekman, P.: An argument for basic emotions. *Cognition and Emotion* 6, 169–200 (1992)
- Fiorelli, M.: *Percezione come Abduzione (Tesi triennale)*. Roma, La Sapienza (2005)
- Hanson, N.R.: *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press (1958)
- Gärdenfors, P.: The dynamics of belief as a basis for logic. *British Journal for the Philosophy of Science* 35, 1–10 (1984)
- Gärdenfors, P.: An epistemic analysis of explanations and causal beliefs. *Topoi* 9, 102–124 (1990)
- Magnani, L.: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Berlin (2009)
- Moriarty, S.E.: Abduction: a theory of visual interpretation. *Communication Theory* 6, 167–187 (1996)
- Ogden, C.K., Richards, I.A.: *The Meaning of Meaning*. Harcourt & Brace, N.Y. (1923)
- Peirce, C.S.: *Collected Papers*. Harvard, U.P. (1935)
- Pizzi, C.: “Since”, “Even If”, “As if”. In: Dalla Chiara, M.L. (ed.) *Italian Studies in the Philosophy of Science*, pp. 73–87. Reidel, Dordrecht (1981)
- Pizzi, C.: Iterated conditionals and causal imputation. In: McNamara, P., Praakken, H. (eds.) *Norms, Logics and Information Systems*, pp. 147–161. IOS Press, Amsterdam (1999)
- Pizzi, C.: Cotenability and the logic of consequential implication. *The Logic Journal IGPL* 12, 561–579 (2004)
- Pizzi, C.: Gestalt effects in abductive and counterfactual inference. *Logic Journal of the IGPL*, 257–270 (2006)
- Pizzi, C.: Abductive inference and iterated conditionals. *SCI*, vol. 64, pp. 365–381 (2007)
- Shackle, G.L.S.: *Decision, Order and Time in Human Affairs*, Cambridge U.P. (1961)
- Shafer, G.: *A Mathematical theory of Evidence*. Princeton, U.P. (1976)
- Thagard, P.: Coherence, Truth, and the Development of Scientific Knowledge. *Philosophy of Science* 73, 28–47 (2006)

# Not by Luck Alone: The Importance of Chance-Seeking and Silent Knowledge in Abductive Cognition

Emanuele Bardone

**Abstract.** In this paper I will focus on luck in abductive cognition and its relationship with chance-seeking and silent knowledge. By definition luck can be neither predicted nor planned, but I will try to show how it can be actively *sought* by seeking those chances maximizing *abducibility*, which will be described as the opportunity of being afforded by lucky events. I will root this ability of seeking chances in evolution, more precisely, in the ability – not entirely unique to our species – of creating powerful cognitive niches, whose construction and modification lead to humans to self-domestication and, in so doing, to introduction of a sense of purposefulness in evolution. Finally, I will introduce the notion of silent knowledge that will be defined as the form of knowledge that emerges along with chance-seeking activities.

## Introduction

Luck permeates human life. That is a simple fact of life. But when it comes to knowledge and discovery, it seems its role is overlooked. In a way knowledge is precisely a response to luck. One should know about the world in order not to be affected by luck – especially if it is bad luck and so causing troubles. So, it sounds counter-intuitive to think that discoveries can be made just by serendipitous events – out of one’s control. In fact, labeling an event as serendipitous or lucky is somehow degrading the effective work the discoverer has done in order to achieve his or her discovery. But is luck really to be intended as a mere accident beyond scientific and theoretical investigation? Is it really degrading the actual cognitive effort one makes?

A growing literature has questioned the idea that luck should be kept outside of rational investigation. Among the many contributions appeared in the last decade,

---

Emanuele Bardone

Institute of Informatics, Tallinn University, Tallinn, Estonia

and

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

e-mail: [bardone@tlu.ee](mailto:bardone@tlu.ee)



Chance Discovery [28, 30, 1, 26, 2] and *epistemic* luck [32] are by far the most interesting contributions on the matter. In this paper I will focus on luck and its relationship with chance-seeking and silent knowledge.<sup>1</sup> More precisely, I will focus on how humans try to get lucky – or eureka, as I will show – by smart eco-cognitive manipulations, and how that has an impact on our evolution. The paper will proceed as follows: in the first part I will illustrate my interpretation of luck and its relevance for abduction cognition, namely, that kind of cognition primarily devoted to the selection and/or creation of hypotheses. My main claim is that luck is cognitively relevant insofar as it contributes to affording human beings to generate or select the correct hypothesis solving a problem one is facing. I will then present the main idea of the paper: by definition luck can be neither predicted nor planned, but it can be actively *sought* by seeking those chances maximizing *abducibility*, which will be described as the opportunity of being afforded by lucky events – events that are out of one’s control.

I will root this ability of seeking chances in evolution, more precisely, in the ability – not entirely unique to our species – of creating powerful cognitive niches, in which the environment serves the purpose of maximizing abducibility. Indeed, evolution – in its Darwinian sense – does not display any purposefulness, as it is characterized by what I call *evolutionary luck*. However the evolution of our species as powerful eco-cognitive engineers constituted a turning point. Even though evolution still remains *blind*, through the construction of powerful cognitive niches, our species has introduced a second inheritance system – an eco-cognitive one – in which potentially benefiting chances for taming or at least lessening the negative impact of evolutionary luck had been uncovered. I will argue that by cognitive niche construction humans come to self-domestication: that is, they partly affect their own evolutionary trajectory.

Self-domestication does not lead to rule out the role of luck in abductive cognition. The last section of the paper aims to shed light on a particular kind of knowledge I will call *silent* knowledge. As I will illustrate, the chance-seeker cannot entirely rely on a well-defined stock of knowledge, which pre-exists the exploration of the environment. Conversely, the chance-seeker leans on silent knowledge, which emerges in the process.

## 1 From Pure Luck to Chance-Seeking

An interesting example – though fictional – illustrates the role that luck can play in abductive cognition. In one of the episodes of the American television medical drama *House MD*, the main character – the diagnostician Dr. Gregory House – is dealing with the case of a 70-year-old scientist who collapsed in his laboratory while doing some experiment on rats. After the usual trial and error process, Dr. House successfully solves the case arriving at the correct diagnosis – amyloidosis, which consists in the abnormal deposition of amyloid, a particular protein, in various

---

<sup>1</sup> A terminological note is required: in the course of the paper I will mainly use the word luck to refer specifically to luck in abductive cognition.

tissues of the body. What is interesting here is the way Dr. House came up with the diagnosis. The story develops as follows. Dr. House is hanging around in the hospital when he stumbles upon an underaged girl he previously met. Her father was cured some days before and now she comes back to the hospital where Dr. House works claiming that his dad lost his medicine and she has to refill. After a short conversation, the girl leaves turning her back to Dr. House who eventually checks out her Congo red thong-covered ass. Some second later Dr. House gets his insight: it might be amyloidosis. How did he come up with the explanation? Amyloidosis can be confirmed by performing a test called Congo Red Dye Test. Congo red is used to stain microscopic preparates. That is, it is added to a sample of patient's tissue and then put under the microscope. Under polarized light it indicates the presence of amyloid fibrils, as the amyloid tissue turns a dark red.

As the example shows, there are several features that we usually attribute to things happening by luck. For instance, the lucky event is *out of control* and *accidental*. It is also *unique* or *singular* in the sense that it is usually perceived not as the result of a process, but as a sudden *Gestalt switch*. A lucky event is *rare* and in a way it may happen to everybody, thus it is *universal*. Finally, it is consequential, meaning that it has some consequence. Now, I claim that a lucky event has two other important features. The first is that it is *eurekaean* and the second is *ecological*. Let us start from the first.

*Eureka* is an exclamation, which in ancient Greek means "I have found it". It is most commonly attributed to the ancient Greek scholar Archimedes who reportedly proclaimed it to celebrate his discovery. *Eurekaean* is nothing but a neologism to indicate that a lucky event has an impact on our ability to come up with a good hypothesis, which turns out to be the solution for the problem at hand. From an abductive point of view, I may argue that the lucky event is eurekaean as far as it makes visible to the abducer the clue (or the set of clues) enabling him or her to infer the correct hypothesis. In our example, the girl's dark red thong made visible one of the crucial symptoms for amyloidosis. This is what Magnani [23] called *abducibility*. In Magnani's view, abducibility is a characteristic of any mediating structure (objects, artifacts, symbols, etc.), which makes available the way an event came about, that is, its past history. For example, the snow displays a high level of abducibility, because people and/or animals leave footprints, from which we can infer where they went and which trajectory they took. In this sense, as Magnani noted, abducibility is recoverability. Meaning that, given a certain event, we can go back in time to those other events that originate it. A quite sophisticated example is provided by psychoanalysis, in which symbols – mainly in the form of verbal explanations – allow us to maximize abducibility as recoverability. As Magnani put it, explanations (but also other artifacts like drawings) in the therapeutic setting "can emerge thanks to the fact that symbols are memory mediators and, moreover, maximize abducibility (recoverability) of their [patients'] past history, that is of all the psychic events that originate them" [23] p. 214]. Symbols as well as artifacts store information of the past, and, as long as they are available, they give us the chance to have access to something that is *no longer* available. I will come to this issue in section 3 where I will discuss the importance of cognitive niche.

As just noted, a lucky event functions as a trigger for the abductive process that leads to the solution of a problem. This process – I posit – is *mediated* by the cognitive niche one lives in. In this respect, a lucky event is ecological in the sense that it may afford us to guess. For luck is not something that we possess, but – as any other resource we happen to find in the environment – something we may be exposed to. In our example, there is no causal connection between Dr. House’s diagnosis and the red thong. Conversely, the red thong on display *afforded* or *suggested* Dr. House to guess the right hypothesis. I will come back to this later.

As already noted, luck cannot be planned anyhow. Indeed, there are certain situations that may be described referring to *pure* luck. To some extent this is the case approximated by what I will call *evolutionary* luck. Even though it represents more an ideal condition than an actual fact, pure luck *just* affects the cognitive agent. I will come back to this in the next section. Pure luck aside, in all other situations the cognitive agent – the human one in our case – is not entirely passive. Humans are afforded in different ways by external circumstances. That is, a lucky event is *eu-rekan*. I have posited that epistemic luck is not neutral in the sense that it prompts a certain reaction. Interestingly, in our example the dark red thong helps Dr. House select the correct disease among all those plausible, because he *could be* afforded by the event. That is, his medical knowledge made possible for him to grab the chance delivered by luck. Clearly, a person who did not know about the Congo red would not have been afforded by the thong. Therefore, luck could have not brought about any substantial effect – whether negative or positive. I will come back to this issue in the last section of the paper when I will introduce the notion of silent knowledge. As I will try to show, silent knowledge is that form of knowledge that benefits from luck, as it can only be recognized or activated as relevant only after a lucky event takes place (i.e., the red thong for Dr. House).

The possibility to be afforded to make an inference is fundamental in order luck to provide us with chances to grab. More than a century ago Louis Pasteur observed that luck favors only the *prepared* mind.<sup>2</sup> It seems that knowledge plays a crucial role for enhancing our ability as chance seekers. Even though luck cannot be controlled or planned anyhow, I posit that it can be taken away from captivity so as to maximize abducibility. That is, the opportunity that is accidental, singular, out of our control may afford/suggest us to select and/or generate the correct hypothesis. Let us make a second example.

In 1990 in Vietnam malnutrition was affecting the majority of children. Dr. Jerry Sternin was sent there on behalf of a NGO called “Save the Children” to try to figure out what to do in order to mitigate such plague.<sup>3</sup> Indeed, he had no chance to tackle down the problem, which was dependent on other broader issues like poverty, ignorance, and poor sanitation. His budget could allow him to do nothing, but one thing: he recruited local mothers to weigh the children in the villages. By doing that he could find out those children who were not underweight and consequently analyze

---

<sup>2</sup> Pasteur’s quote is usually connected with serendipity, which is generally defined as making an accidental discovery. In my view, serendipity may be considered a sort of pure form of luck.

<sup>3</sup> The case is reported in [18].

their family background. What he discovered after that was quite interesting. He identified a group of mothers – not belonging to any of the rich and influential families – who used to give their children a bowl of plain rice like any other mother, but adding shrimp, crabs, and sweet-patato greens. In doing so, their children actually got a daily portion of proteins and so they could stay healthy.

This case is quite different from the previous one. The case could be solved after Sternin identified the group of mothers adding shrimp and crabs to the bowl of plain rice. It was not just *plain* luck. Actually, he acted on the environment by performing a *manipulative abduction* [22, 23]. Generally speaking, manipulative abduction is the process in which a hypothesis is generated and/or selected *through doing*. For example, if we are given a birthday present and we want to know what the box contains without unwrapping it, then we are prompted to shake it. Interestingly, we do not come to shaking it after a process of reasoning from premisses necessarily put into words. Conversely, it is the product of *direct* manipulation of proximal stimuli we receive from the environment. Direct manipulation helps us build up a proto-analysis of the task, which is then the start point for progressively more sophisticated hypothetical explorations of the environment that are both affecting and affected the information we acquire during the process. Indeed, such hypothetical explorations may later involve language-based behaviors along with higher forms of cognition. That view drastically contrasts with the one claiming that motor activity is only initiated at the endpoint of a very complex process in which a detailed representation is created [6].

At a more abstract level, manipulative abduction partly re-conceptualizes the way human cognitive complexity originates. I claim that human cognitive complexity is not viewed as a result of top-down process, but as emerging in a questioning process, with question-answer steps [16], which involve an agent and his or her environment. Within such a framework, manipulative abductions can be considered as environmental interrogations, which are progressively refined as the environment – appropriately modified – is transformed from a source of constraints into a source of resources.<sup>4</sup> I will come back to this issue in sections 3 and 4. In my second example, Sternin immediately operates on the environment in order to overcome the paucity of information and options available to him. The decision of weighing children can be considered the result of a manipulative abduction, which allowed Sternin – in a subsequent chain of abductions partly manipulative and partly not – to generate the hypothesis about the local wisdom of that group of mothers: underweight children's mothers do not add shrimps and crabs to plain rice when they could, if appropriately instructed. Interestingly, Sternin's abductive cognition is still affected by luck, because he could not predict that weighing children would eventually allow him to spot the group of wise mothers, and so solve the puzzle. However, his *manipulative guess* was not a blind one. Conversely, it was performed so as to increase his

---

<sup>4</sup> A description on how a set of constraints can be transformed into a resource is provided in [46].

chances to stumble upon something potentially useful yet unknown for solving his problem.<sup>5</sup>

I may now derive some interesting implications. First of all, luck can be sought by *seeking* all those situations, namely, chances, in which we are potentially afforded to generate the correct hypothesis in concert with environmental resources. Which means that our manipulative guessing aims to uncover *affordances* in the environment that help to solve the problem.<sup>6</sup> Second, the ability of seeking chances rests on our knowledge. In fact, chances are those situations in which abducibility is maximized. But abducibility cannot be maximized, if one lacks knowledge in terms of abductive skills required to be afforded [26]. Thirdly, the kind of knowledge I am talking about is eco-cognitive in its essence: it is the kind of knowledge which enables us to potentially make the best out of the environment by establishing *structural couplings*.

More generally, I claim that maximizing abducibility is carried out at the *eco-cognitive level*. That is, humans maximize their chances to be lucky by constructing cognitive niches so as to be better afforded as abductive agents. This is in line with what I argued above: the maximization of abducibility is always related to a mediating structure, which helps recover the past history of a given event. That is possible, as the mediating structure stores those clues that help us infer how a certain event came about. In the following sections I will illustrate how our my account about luck and chance-seeking might be fruitfully applied to evolution in order to shed light on a quite controversial and hotly debated topic, namely, the role of *purposefulness* in evolution. Neo-darwinism states that there cannot be any role for purposefulness in evolution. Living organisms as part of a species evolve without following a pre-determined path, which makes sense of the various adaptations. We might say that what drives evolution is *evolutionary* luck. That is, organisms happen to develop adaptive solutions for their survival and reproduction simply by (evolutionary) luck. In the following I will show how purposefulness may emerge in evolution in terms of chance-seeking.

## 2 The Notion of Evolutionary Luck

According to the traditional view [27], there are four major features characterizing evolution: *multiplication*, *variation*, *heredity*, and *competition*. Multiplication refers to the fact that an entity can reproduce and in doing so it can give two, three or more others. Variation accounts for the fact that not all entities are identical. Heredity means that different entities will produce different entities. So, for instance, entities of type *A* will produce entities of type *A*, whereas entities of type *B* will produce

<sup>5</sup> Manipulative guessing drastically benefits from the so-called tacit dimension of discovery. For a discussion of this issue with relation to abductive cognition, see Magnani [22] ch. 1]. I will be back to tacit knowledge in section 4.

<sup>6</sup> The notion of affordance was introduced by Gibson, who defined an affordance as “an opportunity for action” [13]. For example, a chair affords sitting, water affords swimming, stairs afford climbing. I provided an abductive account of affordance in [2] ch. 4].

entities of type *B*. The last feature is competition. Competition refers to the fact that a given variation has different consequences in terms of survival and multiplication for the entities that inherited it.

One of the most important issues concerning natural selection deals with variation. In an ideal world the entity with the greatest ability of surviving and reproducing will sooner or later outnumber all others. So, it will be the only existing entity. That is not what actually happens in the real world. In fact, from generation to generation some variation may occur so as to produce a complex functional system. According to the traditionally accepted view of Darwin's theory, variations are random in origin [27]. Although some biologists have recently started to challenge some of the main assumptions behind the idea that variations are random [17], it is worth noting here the role played by evolutionary luck is somehow analogous to the one played by epistemic luck. Let us see how.

As I have already mentioned, an ideal world, in which variations do not occur, would end up after a number of generations with one entity outnumbering all the others. However, that might put life at a great risk. What if after some generation the environmental conditions changed so as to make the only species left in our ideal world unfit to survive? Indeed, life would end. As already mentioned, this is not what actually happens in our real world, and evolutionary luck plays an important role here. Since environmental conditions, namely, selection pressures, do change from time to time, in order life to persist there should be a mechanism, which favors mutation and thus variation, *when it is needed*. I maintain that evolutionary luck is what does that job.

When the rate of environmental change is quite low, mutation is not particularly benefiting. We may argue that genes preserve by reproduction those pieces of information which allowed the organism to successfully solve some problem in a given environment. We might argue that genes store information referring to those situations in which evolutionary luck selected the organism. For organisms have a disposal what Lablonka and Lamb [17] called DNA-care-taker genes. Basically, such genes observe and direct the execution of DNA reproduction. When some errors occur in the copying process, they are promptly fixed.

However, when selection pressures dramatically change, these DNA care-taker genes are turned off and evolutionary luck takes over, as the information stored in genes is out of date, so to say. Indeed, as the rate of mutation increases, so does the possibility for an entity to develop a maladaptive mutation and thus being selected out. In a way, to use an analogy introduced by Lablonka and Lamb [17], it is like for poor people to buy a lottery ticket with the last coin left in their pocket. They might get billionaire or, going to the other extreme, they might not win and so starve and eventually die. Indeed, this last case is extreme, but it makes more visible how mutations and variations are not designed or planned. Meaning that the survival of a given entity does not respond to any particular design – intelligent or not it does not matter. It is just the product of evolutionary luck. More generally, evolutionary luck is responsible for the emergence of a particular adaptation, which actually makes the bearer survive and reproduce at a higher rate. As already stressed above, adaptations are not chosen by the evolving organism, but they just appear.

An interesting case of evolutionary luck is the emergence of a certain trait as a *by-product*. Although it is still highly controversial [33, 38], some researchers refer to religion as a by-product of evolution [5]. There is no specific input-restricted mechanisms or dedicated domains explaining the fact that some people develop various beliefs in supernatural agents [33], which are supernatural in the sense that they explicitly violate some of the most elementary laws of folk physics. Actually, what is for some people like an instinct – the religious one – is resulting from the evolution of other adaptations, namely, those cognitive mechanisms cognitive enabling humans to reason about the intentional states of others, where the word “others” includes all types of agents, for instance, absent or dead persons, fictional characters, and so even supernatural beings. So, there is no evolutionary cause as to why some people believe in God. It was not an ability that was selected *per se* during human evolution. It was just by luck that our ancestors evolved specific beliefs related to supernatural agents as resulting from what Richard Dawkins called “misfiring” [9]. To be precise, Dawkins mainly referred to the misfiring of kin altruism genes for explaining the emergence of human morality. However, the same can be said about religious beliefs: mind-reading got wired into the human brain for a specific purpose: to read the intentional states of others. The misfiring of mind-reading related mechanisms made possible the emergence and selection of religion as a *natural phenomenon*, that is, the extension of anthropomorphic thinking beyond living creatures. What is worth noting that this misfiring Dawkins talks about may also be described in terms of chance-dynamics like in our second example. That is, mind-reading related mechanisms did not *determine* the emergence of supernatural beliefs, but they gave the chance for supernatural beliefs to emerge and flourish in the *milieu* of the human mind.<sup>7</sup> That is, mind-reading related mechanisms provided our ancestors with the chance to think of natural forces as *kind of agents*, which therefore share some features with human beings and animals. I may even say that such mechanisms afforded our ancestors to extend the application of the idea of agency.

In sum, evolution is not designed to achieve a particular outcome. Foresights are not possible to make. In the following I will show how humans try to tame natural selection by means of cognitive niche construction. That is, the construction of more and more sophisticated cognitive niches, which may enhance humans’ chance of being afforded by luck. The illustration of cognitive niche construction will also shed light on the issue of purposefulness in evolution.

### 3 Artificial Selection, Niche Construction, and Chance-Seeking

In the discovery and exposition of natural selection the analogy with artificial selection by breeding, namely, domestication, seems to have played a key role in Darwin’s line of reasoning [40]. Darwin noted that many species have been modified by breeders, for example, animals (cattle, sheep, dogs), flowers, and vegetables. In the

<sup>7</sup> An abductive account about how supernatural beliefs are generated is given by Bertolotti and Magnani [4].

analogy with breeders, Darwin conjectured that something similar could be done by the *impersonal* environment via *natural* selection with no intervention whatsoever [8], although he could not provide a single instance of speciation by natural selection [11]. For design revision could be explained as gradual change leading to speciation, and so ruling out the activity of any creator or designer. The differences between artificial and natural selection is my focus here, because it implicitly addresses the problem of evolutionary luck and purposefulness in evolution. Let us see how.

In natural selection the selector (say, Mother Nature) is not the one who benefits from selection. The one which benefits from selection is the selected. In this respect evolution by natural selection is contingent and does not look forward: it is goalless and purposeless. This is in line with what I have argued above: evolutionary luck may benefit a species rather than others, but its “reasons” are concealed or, more simply, they are unintelligible. In the case of variation under domestication, namely, artificial selection, things are different. The one which benefits is not the selected, but the selector, namely, the human domesticator. Interestingly, Darwin pointed out that “nature gives successive variations; man adds them up in certain directions useful to him” [8, p. 22]. So, certain traits selected in domesticated animals benefit not the animal *per se* but the domesticator, which is said to be the major source of selection. Darwin also claimed that deliberate choice made by the selector is not essential. In his own words: “a man who intends keeping pointers naturally tries to get as good dogs as he can, and afterwards breeds from his own best dogs, but he has no wish or expectation of permanently altering the breed” [8, p. 25]. That is what Darwin called “unconscious selection”. So, by domestication humans try to more or less tacitly exploit evolutionary luck so that certain evolutionary trajectories may serve human’s purpose, not nature’s one, if any. It is worth noting that Darwin also mentioned what he called “methodical selection”. Methodical selection is characterized by the selection of a new strain or sub-breed so that it is superior to any others. According to Darwin, this second type of artificial selection emerges only after unconscious forms, that is, when a given breed is already improved and modified without any methodical intervention. As Darwin put it:

[s]low and insensible changes of this kind could never be recognised unless actual measurements or careful drawings of the breeds in question had been made long ago, which might serve for comparison. [8, pp. 25]

Humans try to take advantage as much as possible of variations already selected via natural selection and re-direct them according to their needs. In so doing they alter the breed. Now, I may argue that domestication may be considered one of the possible ways in which we try to *domesticate* evolutionary luck. In a way, it may be considered as one of the several instances of “self-domestication”. By self-domestication I refer to that particular activity or set of activities in which humans manipulate and modify the environment – and everything in it, including plants, animals, etc. – so as to increase *their own* chances of survival and reproduction. In order to spell out my proposal, I briefly present the main tenets of the theory of niche construction.



The theory of niche construction has been recently advocated by a number of biologists, who revisited some basic aspects concerning Darwinism in order to acknowledge the agency of living organisms and its impact on evolution. It is crucial to this theory the revision of the notion of *adaptation*.

Generally speaking, evolution is a response to changes in selection pressures all living organisms are subjected to [19]. Traditionally, the environment has been always meant to be the only source of selection pressures so that adaptation is considered a sort of top-down process that goes from the environment to the living creature [14]. In contrast to that, a bunch of biologists [20, 10, 29] has tried to provide an alternative theoretical framework by emphasizing the active role played by the living organisms.

We may say that the environment is meant to be a sort of “global market” that provides creatures with unlimited possibilities. Indeed, not all the possibilities that the environment offers can be exploited by the organism. For instance, the environment provides water to swim in, air to fly in, flat surfaces to walk on, and so on. However, no creatures can be afforded by all of chances. If it were not like that, we would have no evolutionary luck at all. But what is important to stress here is that all organisms try to modify their surroundings in order to better exploit those elements that suit them and eliminate or mitigate the effect of the negative ones.

According to this view, adaptation is meant to be *two way*. Organisms (humans, in our case) adapt to their environment *and vice-versa* [19]. That is, environments cause some features of living creatures to change. But also creatures cause some features of the environments to change. In altering their local environments, living creatures contribute to construct and modify the so-called ecological niches. As Reed added:

[...] only the relative availability (or nonavailability) of affordances create selection pressure on the behavior of individual organisms; hence behavior is regulated with respect to the affordances of the environment for a given animal<sup>8</sup> [35, p. 18]

In any ecological niche, the selective pressures of the local environment are modified by organisms in order to lessen the negative impacts of all those elements which they are not suited to. Indeed, this does not mean that natural selection is somehow halted. Rather, this means adaptation cannot be considered only by referring to the agency of the environment, but also to that of the organism acting on it. In this sense, animals and other living creatures are ecological engineers, because they do not simply live their environment, but they actively shape and change it [10].

In case of humans, the ubiquitous presence of *cognitive* niches contribute to introducing a second and non-genetic inheritance system insofar as the modifications brought about on the environment persist, and so are passed on from generation to generation [29]. For humans, which had become extremely successful eco-cognitive engineers, the main advantage of having a second inheritance system is that it enabled them to access a great variety of information and resources never personally experienced, resulting from the activity of previous generations [29]. That is, the information and knowledge humans can draw on are not simply transmitted, but they

---

<sup>8</sup> A discussion on Reed’s stance on evolution and affordance can be found in [45].

can also be accumulated in the so-called cognitive niches. Indeed, the knowledge we are talking about embraces a great variety of resources including knowledge about nature, social organization, technology, the human body, and so on. In any cognitive niche the relevant aspects of the local environment are appropriately selected so as to turn the surroundings – inert from a cognitive point of view – into a mediating structure delivering suitable chances for behavior control [29].

Ecological inheritance system is different from the genetic one in the following way [29]: 1) genetic materials can be inherited only from parents or relatives. Conversely, modifications on the environment can affect everyone, no matter who he/she is. It may regard unrelated organisms also belonging to other species. There are several global phenomena such as climate change that regard human beings, but also the entire ecosystem; 2) genes transmission is a one way transmission flow, from parents to offspring, whereas environmental information can travel backward affecting several generations. Pollution, for instance, affects young as well as old people; 3) genetic inheritance can happen once during one's life, at the time of reproductive phase. In contrast, ecological information can be transferred during the entire duration of life. Indeed, it depends on the eco-engineering capacities at play; 4) genetic inheritance system leans on the presence of replicators, whereas the ecological inheritance system leans on the persistence of whatsoever changes made upon the environment.

Coming back to artificial selection and self-domestication, my take is that humans domesticate themselves by domesticating the environment. That is, self-domestication relies on a fundamental circularity, in which humans domesticate the environment – meaning that they modify it via niche construction – and in so doing they are domesticated. In this regard I may argue that human niche construction can be considered as a peculiar form of artificial selection, which is characterized by the identity of the selector and the selected (or the domesticator and the domesticated). From this identity I may derive an interesting conclusion. The main product of self-domestication is the selection of *epigenetic openness*. Epigenetic openness – appropriately favored at genetic level – promotes the generation/selection of more *plastic* adaptive solutions, which rely on “domain-specific learning in the semiotic biocultural complex, in particular language” [37]. In my view, that is nothing but the development of more and more sophisticated activities of chance-seeking along with the active modification of the environment.

The activities of chance-seeking drastically benefit from the high level of *plasticity* which is exhibited by humans as a product of self-domestication. I posit that plasticity helps humans exploit latent chances and enhance (the maximization of) abducibility. Let us see how. Plasticity of response to an ever-changing environment is connected to the necessity of having other means for acquiring information, more readily and quickly than the genetic one [15]. I posit that (cognitive) niche construction plays a fundamental role to meet this requirement. Plasticity depends on cognitive niche construction as far as humans may create and store profitable structural couplings with their local environment, which may be later modified and improved, if necessary. This establishes a sort of loop, in which the activity of cognitive niche construction liberates additional cognitive chances for behavior, which

may later be exploited and directed for improving pre-existing cognitive niches. As Magnani noted “the mind grows up together with the representational delegations to that ‘nature’ that the mind itself has made throughout the history of culture by constructing the cognitive niches” [25]. In sum, cognitive niches are crucial in developing more and more sophisticated forms of plasticity – and thus chance-seeking activities – because cognitive niches constitute a fundamental source of information and cognitive resources favoring the maximization of abducibility and thus the domestication of luck. That introduces “a sense of purposefulness” in evolution that is worth investigating.

As mentioned above, variations by natural selection – which are ultimately chances for a species to survive and reproduce – are delivered by evolutionary luck. As Darwin pointed out, nature benefits the selected without the selected can control the process. But by niche construction humans are able to *re-direct* their evolutionary trajectory so as to benefit themselves.<sup>9</sup> Here I can paraphrase what Darwin said about artificial selection: evolutionary luck as natural selection brings about successive variations; humans add them up in certain directions useful to them via niche construction. As already mentioned, the main product is epigenetic openness and the emergence of purposefulness in evolution.

At a less abstract level, the sense of purposefulness is stressed by the fact that the modifications made on the environment via niche construction are not random or entirely due to luck. Humans seek and select appropriate chances that are the result of *abductive* manipulations of the environment [26]. Besides, insofar as benefiting modifications persist in the cognitive niche, they enter the scene as chances already available *over there* so that evolutionary luck is not the only factor affecting evolution.<sup>10</sup> In this sense, as argued by Turner, “evolution becomes less a province of one class of arbiters of future function – genes – and more the result of a nuanced interplay between the multifarious specifiers of future function” [41, pp. 348–349]. Now, that a sense of purposefulness appears in evolution means that in a way human beings are in a better position for increasing their fitness. This is implicit in the definition of niche construction: human beings are not merely passive, but they actively seek out chances for reproduction and survival by constructing *cognitive niches*.

Having clarified the evolutionary process leading from evolutionary luck to chance-seeking, let us go back to the second example introduced in section 1. I claimed that Sternin was indeed helped by luck to find the bright spot – the group of wise mothers. However, putting the children on a scale was not the result of a blind guess, but a hypothetical manipulation of the environment, which was primarily meant for exploring chances that, in turn, helped him uncover further opportunities for action, namely, affordances. So, in a way I may say that it was *by luck* that

<sup>9</sup> It is worth noting that cognitive niches may bring about maladaptive consequences, as natural selection is not halted by niche construction. To face maladaptive consequences humans may create counteractive niches [29], which are precisely meant to lessen such human-induced negative impacts. A theoretical account about how technology may promote the construction of maladaptive cognitive niches is provided in [3].

<sup>10</sup> For a discussion of the moral implications of cognitive niche construction, see Magnani [24].

Sternin bumped into the group of wise mothers. But, at the same time, it was also his ability to make the best out of what he had at disposal that helped him solve the case. I may say that Sternin's behavior is regulated with respect to the affordances he finds in his cognitive niche. In fact, Sternin's ability to seek out the best chances cannot be considered in isolation, that is, without the reference to the cognitive niche he happened to be in. First of all, Sternin could rely on a piece of equipment available in the local cognitive niche, namely, a scale. A scale is not just over there in the natural world. But it is something designed and used by humans for a specific purpose. It provides a kind of data that one would not be able to obtain otherwise. More specifically, the scale – as part of Sternin's cognitive niche – gave him access to clues, which afforded him to assess for malnutrition. As a matter of fact, without the scale that would be simply out of his reach. Interestingly, there is no causal link between having a scale and weighing the children. But the former *affords* or *suggests* the latter.

Secondly, the data gathered thanks to the scale provided additional chances for action. In fact, Sternin did not immediately spot the group of wise mothers. Before doing that he *collected* further data and information about the family background of each child so as to filter out those who had no problem just because their family could afford a richer but more expensive diet. Here again there is no causal link between the children's weights *and* the idea of collecting information related to the family background of each child. But the former afforded Sternin to do the latter.

There is a third point to stress. Sternin was a doctor. That means that he had medical knowledge, which was fundamental in many respects. For instance, it was essential for the identification of malnutrition, which is a medical diagnosis: being skinny is not enough to say that somebody has a problem. Sternin could rely on decades of scientific research about all nutrients (carbohydrates, fat, protein, minerals, etc.) and their importance for a balanced diet. More in general, he could benefit from a powerful cognitive niche comprising agencies of various kinds (institutional, cultural, human, technological), which ultimately made possible generation by generation the creation, accumulation, and preservation of medical knowledge that is now available to him. In this sense, medical knowledge is a sort of eco-cognitive inheritance system that is delivered via cognitive niche and drastically empowers – where it is at hand – human collectivities to face various problems related to health and well-being.

#### **4 Chance-Seeking, Silent Knowledge, and Luck**

So far I have tried to point out that luck cannot be controlled or predicted, but by seeking good chances problem-solvers can try to get lucky (or eureka, as I argued). I have posited that chance-seeking is manipulative in its essence, as it leans on the detection and exploitation of opportunities for action, namely, affordances, which are available in one's cognitive niche. The development of more and more sophisticated cognitive niches empowers human collectivities to maximize abducibility. It is worth noting that cognitive niche construction is still a hypothetical activity,

meaning that it mostly leans on this manipulative guessing I talked about in section 1. Therefore, luck is still affecting our abilities as chance-seekers. This last section is precisely devoted to showing this sort of interplay between our knowledge and luck within the framework of chance-seeking.

I posited that affording should not be mistaken for causing. By “causing” I mean the production of an effect regardless agent’s intention [35, 44]. For example, a chair affords sitting. That does not mean a chair *causes* a person to sit. I can see a chair and consequently being afforded to sit down on it, but I can decide not to do that. One may claim that to pick up an affordance implies to have a purpose or intention. After hours spent on shopping one may eagerly look for a chair. In this case, we already *know* what to look for in order to be afforded. Interestingly, Reed defines intentions not as causes of action, but *patterns of organization* of action, which are embodied in performances [34, p. 62]. On a more philosophical level, I may claim that intentions are primarily *individuating*. That is, they organize experience as *my* experience. That gives nothing but *direction* to my engagement in the environment. In this regard, I agree with Laurent and Ripoll who claim that cognition is not to be understood as “an interpretation tool, but rather as a directional force” [21, p. 139].

However, there are situations in which I clearly do not know what to look for. In other words, I do not know how to act on the environment so as to be sure that I will find good chances. In a way, this echoes the so-called Meno’s Paradox which was illustrating by Michael Polanyi as follows: “[. . .] if you know what you are looking for, then there is no problem. If you do not know what you are looking for, then you cannot expect to find anything” [31, p. 22] <sup>11</sup>

The same can be said about good chances: if you know what good chances are, then they are no chances at all. If you do not know them, you cannot expect to find them. Going back again to our second example, Sternin did not and could not know if weighing the children may help him find a solution for malnutrition. That was a mere hypothetical manipulation, namely, a manipulative abduction, which *only later* turned out to be a good chance.

I have illustrated the importance of manipulative abduction in the previous section. I argued that manipulative abduction is a form of hypothetical reasoning that is performed *through doing*. *Through doing* means that knowledge guides action, but also vice-versa in a sort of feedback loop. That means that action is not called at the endpoint of a process in which a rich representation is built up. Conversely, it already takes part in an early stage providing both input and output for successive and more sophisticated cognitive explorations of the environment. So, in a way the kind of knowledge that is actually employed in chance-seeking activities is not and cannot be *defined* beforehand. But it is somehow *enriching* as well as *enriched* by. In the rest of this section I will try to better characterize such a kind of knowledge.

---

<sup>11</sup> Simon provided a solution of the paradox by referring to the so-called “generate and test” method. Basically, he argued that “our ability to know what we are looking for does not depend upon our having an effective procedure for finding it: we need only an effective procedure for testing candidates”. [36, p. 339].

During his 2005 commencement address at Stanford the American Entrepreneur Steve Jobs proudly told an interesting story<sup>[12]</sup> which is worth recalling here. After one semester at Reed College, Jobs decided to quit. As a college dropout he did not have to take the normal classes. So, he decided to take a calligraphy class – the best calligraphy class in the country, as he later recalled. What he learnt during this class remained basically unused for about ten years, until Apple started to design the Macintosh computer. It was only at that time that he decided to design beautiful typography all into the Mac, and thus exploit his knowledge in the field. He proudly concluded his story saying that no personal computer would have multiple typefaces or proportionally spaced fonts, if he as a college dropout had never taken a calligraphy class.

Whether Jobs' claim is true or not, it does not matter. As far as I am concerned here his story shows how discovery and innovation can hardly be reduced to a *linear* collection of events, in which all the dots are clearly connected with each other *beforehand*. This has an important theoretical consequence to draw with relation to knowledge, chance-seeking, and luck. Our pieces of knowledge enabling us to successfully solve a problem or make a discovery may come from different, multiple, asynchronous, and unsorted sources.<sup>[13]</sup> Those pieces of knowledge might go *silent* for quite a long time before they can profitably be used and have unity. For it seems that there is a particular kind of knowledge – I may call *silent knowledge* – which becomes crucial for seeking out good but unexpected chances. By *silent* I refer to the fact that some piece of knowledge may be removed from sight, but still *there*.

Silent knowledge is not to be taken as necessarily tacit. Tacit knowledge was introduced by Polanyi to refer to the fact that we may know how to do something, say, swimming, without being able to explain/express *that* which we know. Our knowledge comprises also all those skills that are embodied – performed by the body. So, the term “tacit” refers to the fact that our knowledge to be as such does not need to be fully expressed in *words*. Let us say that it is also *implicit*. In the case of silent knowledge, the word “silent” refers to the fact that some pieces of knowledge remain *concealed* or *inactive* for much of the time so that they cannot be used. Yet they are *over there*. Going back to Jobs' story, his knowledge about calligraphy and typography is not tacit, but for years *he could not see how to use it*. That is what characterizes silent knowledge.

Silent knowledge has an important feature to mention here. It is *narrative*.<sup>[14]</sup> According to Danto, narrative sentences are those which describe something that happened at a particular point in time but, in doing so, they refer to knowledge of a *later* point<sup>[7]</sup>. That is, narrative sentences assume to have a sort of *historical* perspective on a given event. So, for instance, if I say that “then the Hundred years' war began”, I clearly have to know what happened *at the end* of the story, namely, that the war

<sup>12</sup> The prepared text of the commencement address can be found at <http://news.stanford.edu/news/2005/june15/jobs-061505.html>. The story is analyzed in detail by Jay Elliot<sup>[12]</sup>.

<sup>13</sup> I want to stress that silent knowledge can be related to *know-how* as well as *know-that*.

<sup>14</sup> The word “narrative” does not necessarily imply that silent knowledge is then exclusively discursive or sentential knowledge.

which started in 1337 lasted for about a hundred years. The same can be said about almost any statement made in a history book. Now, Danto makes an interesting point about relevance. As he put it:

It will then not be enough simply to be able to predict future events. It will be necessary to know *which* future events are relevant, and this requires predicting the *interests* of future historians [emphasis in the original]. [7] p. 169]

For silent knowledge I may say something similar. That is, its relevance can only be identified after a certain event has taken place. Jobs' case is paradigmatic: the calligraphy class became relevant only later.

The question about relevancy and silent knowledge has an interesting connection with a term which Taleb has recently coined: antifragility [39]. Antifragility is a neologism to refer to those systems that are not fragile but in a specific sense. That is, systems which are not simply immune to randomness, volatility, and uncertainty (that is what Taleb calls "robustness"); but systems which actually *benefit* from them. To use a metaphor, if something fragile carries the label "please, handle with care", something that is antifragile carries the label "please, mishandle". The former gets broken, if mishandled; the latter gets *stronger*. Silent knowledge is antifragile in a rather specific sense: it takes advantage of irrelevancy. Already Sun Tzu acknowledged that the pair relevancy/irrelevancy is not necessary a dialectic opposition. In his words: "[i]f we do not know what we need to know, then everything looks like important" [42]. When one's task is clearly defined, the use of heuristic search methods inevitably filters out information and even piece of knowledge that might turn out to be useful *afterwards*. Conversely, if the epistemic task at hand – for instance, explaining a certain phenomenon – changes or is later re-defined due to its non-monotonicity (i.e., we simply get more information and data), silent knowledge acquires a heuristic value. In this sense, silent knowledge is not only immune to situations in which relevance cannot be clearly distinguished from what is irrelevant, but it also benefits from them. In this regard, luck plays an important role. Let us see how.

On those occasions in which the distinction between what is relevant and what is not is clear-cut, our beliefs becomes less immune to all that information which falls off our focus. That is, we may fail to notice what might later become the major source of inconsistencies or falsifications as our epistemic tasks proceed. Inconstancies and falsifications are unexpected events, which emerge by luck in the course of one's investigation. Conversely, silent knowledge survives inconsistencies, and luck is precisely what may help us find out new and meaningful connections. This may clarify Sun Tzu's argument based on what may be called *transparent* relevance: when we do not know what we need to know, luck may show to us something potentially relevant, namely, good chances. It follows that silent knowledge exhibits its values especially when the impact of luck is significative. Or, better, luck *detonates* it.

There is one more thing to add. I have argued that chance-seeking implies a sort of active attitude towards luck. As repeatedly mentioned, that does not mean that we can predict or plan *when* luck will befall us. In this section I have clearly pointed out that luck still plays a role, as seeking-chance drastically leans on silent knowledge,

whose significance and unity are immanent to and bound with tentative explorations of the environment. Therefore, chance-seeking – along with silent knowledge – is closely related to what Varela and colleagues called “an embodied (mindful), open-ended reflection” [43, p. 27], whose main purpose is “open-ended examination of experience” [43, p. 85]. That means that chance-seekers do not organize and filter their experience according to some kind of knowledge, which pre-exists action and eco-cognitive manipulation. Conversely, chance-seekers organize their experience by making use of silent knowledge that is therefore *enacted*. By *enacted* I mean that it is created through actions and manipulations, which refer to how the chance-seeker acts. It is worth to note that such actions and manipulations do not come about in cognitive vacuum, as already pointed out in this paper. But they are *already* dependent on pre-existing *structural coupling*, which can be viewed as part of the eco-cognitive inheritance I described in section 2.

## Conclusions

In this paper I have tried to analyze the role of luck in abductive cognition. First of all, I outlined the main characteristics of luck. I argued that luck cannot be predicted, since it is out of one’s control. However, it can be domesticated somehow by maximizing abducibility. That is, one may try to get lucky (or eureka) by manipulating the environment so as to be exposed to profitable chances. I maintained that such eco-cognitive manipulations are crucial also from an evolutionary point of view. Although there is no sense of purposefulness in evolution, through the construction of more and more sophisticated cognitive niches, humans had tried to partially domesticate evolutionary luck by means of self-domestication. That is, they have tried to better control some evolutionary forces affecting their survival as a species by re-engineering the relations with their environment. In the last part of the paper I introduced the notion of silent knowledge. I argued that chance-seeking activities are primarily carried out *through doing*. That is, they do not lean upon a pre-defined stock of knowledge, but on a form of knowledge, which is narrative in its essence. That is, it cannot be immediately recognized as such, but it is enacted by further eco-cognitive explorations.

**Acknowledgements.** This research was supported by the Estonian Science Foundation and co-funded by the European Union through Marie Curie Actions.

## References

1. Abe, A.: Cognitive Chance Discovery. In: Stephanidis, C. (ed.) UAHCI 2009, Part I, HCII 2009. LNCS, vol. 5614, pp. 315–323. Springer, Heidelberg (2009)
2. Bardone, E.: Seeking Chances. From Biased Rationality to Distributed Cognition. Springer, Heidelberg (2011)
3. Bardone, E.: Unintended affordances as violent mediators. maladaptive affects of technologically enriched cognitive niches. *International Journal of Technoethics* 2(4), 37–52 (2011)



4. Bertolotti, T., Magnani, L.: The Role of Agency Detection in the Invention of Supernatural Beings. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 239–262. Springer, Heidelberg (2010)
5. Boyer, P.: Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences* 7(3), 119–124 (2010)
6. Clark, A.: Where brain, body, and world collide. *Journal of Cognitive Systems Research* 1, 5–17 (1999)
7. Danto, A.N.: Narrative sentences. *History and Theory* 2(2), 146–179 (1962)
8. Darwin, C.: *The Origin of Species: By Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life*. Cambridge University Press, Cambridge (2009)
9. Dawkins, R.: *The God Delusion*. Bantam Press, London (2006)
10. Day, R.L., Laland, K., Odling-Smee, J.: Rethinking adaptation. The niche-construction perspective. *Perspectives in Biology and Medicine* 46(1), 80–95 (2003)
11. Dennett, D.: *The Darwin's Dangerous Idea*. Simon & Schuster, New York (1996)
12. Elliot, J.: *The Steve Jobs Way: iLeadership for a New Generation*. Vanguard Press, New York (2011)
13. Gibson, J.J.: *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston (1979)
14. Godfrey-Smith, P.: *Complexity and the Function of Mind in Nature*. Cambridge University Press, Cambridge (1998)
15. Godfrey-Smith, P.: Environmental complexity and the evolution of cognition. In: Sternberg, R., Kaufman, K. (eds.) *The Evolution of Intelligence*, pp. 233–249. Lawrence Erlbaum Associates, Mahwah (2002)
16. Hintikka, J.: *Socratic Epistemology: Explorations of Knowledge-Seeking by Questioning*. Cambridge University Press, Cambridge (1993)
17. Jablonka, E., Lamb, M.J.: *Evolution in Four Dimensions. Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. The MIT Press, Cambridge (2005)
18. Kawasaki, G.: *Enchantment: The Art of Changing Hearts, Minds, and Actions*. Penguin & Portfolio, New York (2011)
19. Laland, K., Brown, G.: Niche construction, human behavior, and the adaptive-lag hypothesis. *Evolutionary Anthropology* 15, 95–104 (2006)
20. Laland, K.N., Odling-Smee, J., Feldman, M.W.: Niche construction, biological evolution and cultural change. *Behavioral and Brain Sciences* 23(1), 131–175 (2000)
21. Laurent, E., Ripoll, H.: Extending the rather unnoticed gibsonian view that? perception is cognitive?: Development of the enactive approach to perceptual-cognitive expertise. In: Araujo, D., Ripoll, H., Raab, M. (eds.) *Perspectives on Cognition and Action in Sport*, pp. 133–146. Nova Publishers, New York (2009)
22. Magnani, L.: *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic/Plenum Publishers, New York (2001)
23. Magnani, L.: *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Heidelberg (2009)
24. Magnani, L.: *Understanding Violence. The Intertwining of Morality, Religion and Violence: A Philosophical Stance*. Springer, Heidelberg (2011)
25. Magnani, L.: Scientific Models are Not Fictions Model-Based Science as Epistemic Warfare. In: Magnani, L., Li, L. (eds.) *Philosophy and Cognitive Science*. SAPERE, vol. 2, pp. 1–38. Springer, Heidelberg (2012)
26. Magnani, L., Bardone, E.: Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction. In: Iwata, S., Ohsawa, Y., Tsumoto, S., Zhong, N., Shi, Y., Magnani, L. (eds.) *Communications and Discoveries from Multidisciplinary Data*, pp. 3–40. Springer, Berlin (2008)

27. Maynard-Smith, J., Szathmari, E.: *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford University Press, Oxford (2000)
28. McBurney, P., Parsons, S.: *Chance Discovery Using Dialectical Argumentation*. In: Terano, T., Nishida, T., Namatame, A., Tsumoto, Y., Ohsawa, S., Washio, T. (eds.) *JSAI 2001 Workshops. LNCS(LNAI)*, vol. 2253, pp. 414–424. Springer, Berlin (2001)
29. Odling-Smee, F., Laland, K., Feldman, M.: *Niche Construction. A Neglected Process in Evolution*. Princeton University Press, New York (2003)
30. Ohsawa, Y., McBurney, P. (eds.): *Chance Discovery*. Springer, Berlin (2003)
31. Polanyi, M.: *The Tacit Dimension*. Routledge & Kegan Paul, London (1966)
32. Pritchard, D.: Epistemic luck. *Journal of Philosophical Research* 29, 193–222 (2004)
33. Pyysiäinen, I., Hauser, M.: The origins of religion: evolved adaptation or by-product? *Trends in Cognitive Sciences* 14(3), 104–109 (2010)
34. Reed, E.S.: The intention to use a specific affordance: a framework for psychology. In: Wozniak, R., Fisscher, K.K. (eds.) *Development in Context: Acting and Thinking in Specific Environments*, pp. 45–75. Lawrence Erlbaum Associates, Hillsdale (1993)
35. Reed, E.S.: *Encountering the World*. Erlbaum, New York (1996)
36. Simon, H.: *Models of Discovery and Other Topics in the Methods of Science*. Reidel, Dordrecht (1977)
37. Sinha, C.: Epigenetics, semiotics, and the mysteries of the organism. *Biological Theory* 1(2), 112–115 (2006)
38. Sosis, R.: The adaptationist-byproduct debate on the evolution of religion: Five misunderstandings of the adaptationist program. *Journal of Cognition and Culture* 9, 315–332 (2009)
39. Taleb, N.: *Antifragility*. Penguin, London (forthcoming, 2012)
40. Thagard, P.: *Computational Philosophy of Science*. MIT Press, Cambridge (1993)
41. Turner, J.S.: Extended phenotypes and extended organisms. *Biology and Philosophy* 19, 327–352 (2004)
42. Tzu, S.: *The Art of War*. Fastpencil, London (2010)
43. Varela, F., Thompson, E., Rosch, E.: *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge (1999)
44. Withagen, R., de Poela, H.J., Araujob, D., Peppinga, J.J.: Affordances can invite behavior: Reconsidering the relationship between affordances and agency. *New Ideas in Psychology* 30(2), 250–258 (2012)
45. Withagen, R., van Wermeskerken, M.: The role of affordances in the evolutionary process reconsidered: a niche construction perspective. *Theory & Psychology* 20, 489–510 (2010)
46. Zhang, J., Patel, V.L.: Distributed cognition, representation, and affordance. *Cognition & Pragmatics* 14(2), 333–341 (2006)

# Cognitive Abduction and the Study of Visual Culture

María G. Navarro and Noemi de Haro García

**Abstract.** In this paper art history and visual studies, the disciplines that study visual culture, are presented as a field whose conjectural paradigm can be used to understand the epistemic problems associated with abduction. In order to do so, significant statements, concepts and arguments from the work of several specialists in this field have been highlighted. Their analysis shows the fruitfulness and potential for understanding the study of visual culture as a field that is interwoven with the assumptions of abductive cognition.

## 1 Introduction

Divergence and consensus are constants in the study of abduction. There are divergences in the exact meaning of the term, but a great consensus on the strong connection of abduction with many disciplines. Magnani [29] has justified and documented all kinds of evidence about the relationship between abductive reasoning and disciplines such as philosophy, legal reasoning, Artificial Intelligence, cognitive sciences, narrative reasoning, decision making, emotional cognition, etc. It is thus reasonable to assume that if abduction is so important as an interplay between this and many other fields it is because cognition in all living beings manifests a clear abductive mark. The idea of ‘abductive cognition’ has been shown to be important thanks to the contribution of Magnani to the vast and complex history of abduction studies.

---

María G. Navarro

Department of Speech Communication, Argumentation Theory and Rhetoric,  
University of Amsterdam, The Netherlands

e-mail: [maria.navarro@cchs.csic.es](mailto:maria.navarro@cchs.csic.es)

Noemi de Haro García

Departamento de Historia y Teoría del Arte, Universidad Autónoma de Madrid,  
Spain

e-mail: [noemi.deharo@uam.es](mailto:noemi.deharo@uam.es)

Walton [47] affirms that abduction is a process of hypothesis formation that is used at the discovery stage of scientific investigation, but we think that it is also a source for a better understanding of both the theoretical and practical dimensions in the study of the humanities. Here we will analyse the presence of abductive cognition in a field of the humanities that has not been explored sufficiently: the disciplines that study visual culture. These are art history and visual studies. It can be said that, in the long tradition of art history, research has been centred on some cultural objects, including some objects of visual culture, that have been selected according to aesthetic criteria. As Dikovitskaya [9] has shown, the ‘cultural turn’ has provoked important changes in the study of the visual such as the marriage between art history and cultural studies that has led to the appearance of visual studies. The interdisciplinary field of visual studies examines the role of all images in culture, trying to go beyond the limitations imposed by aesthetic criteria on the object of the discipline of art history as researchers like Vega [43] have stressed, and claiming that the study of the experience of the visual has to be contextual, ideological and political. Thus visual culture is (in whole or in part) the object of study both of art history and visual studies. Therefore in order to analyse the reasoning process used to think about this object, both disciplines are to be taken into account. As we will show, the abductive reasoning model, which can be more clearly identified in the work of some contemporary specialists in visual studies, is also present in the research of the founders of art history.

The purpose of this article is to explain why research in visual studies must be taken into account in studies in abduction. In order to sustain our thesis, we will establish several conceptual analogies to link both research fields. This will shed a new light on both, and show that abduction is one of the principal characters in the study of the visual. A broad set of concepts could potentially be used to do this, but we will focus on:

1. Conjectural paradigm and re-creative synthesis / inference to the best explanation, helicoid abductive reasoning.
2. Empathy, *pathosformel*, empathetic response / embodiment.
3. The combination of theoretical and manipulative abduction in the study of the visual.

From this analogical reasoning, three consequences are to follow: the first presents abduction as the logical pattern inherent in interpretation. The second is related to perception understood as a limited process. The activity of interpretation can be presented both as a process and as the result of a process where abduction is constantly present. It may appear either as theoretical abduction, as model-based abduction or even as manipulative abduction. The third has to do with the inferential structure of perceived objects. The use of abductive reasoning understood as epistemic change, models the incorporation of new beliefs. The interpretative process and product, the bodily involvement in visual culture experience and even visual culture itself, can be

understood as products that have an inferential structure or that even imply an inferential play.

## 2 Conjectural Paradigm and Re-creative Synthesis

The presence of abductive reasoning in scientific practices related to the arts has been identified in studies that were oriented towards the establishment of a relationship between the interpretation of the arts and semiotics. Ginzburg [16] included the method of the *connoisseur* Giovanni Morelli along with those of Freud and Sherlock Holmes (or better, Conan Doyle's method) in his essays about how in the late 19<sup>th</sup> century a theoretical model for the construction of knowledge emerged in the sphere of the social sciences, the conjectural paradigm. The methods of Morelli, Freud and Doyle had something in common: they were based on taking marginal, irrelevant details as revealing clues to forge their conclusions, and they shared the model of medical semiotics or symptomatology. But the roots of the 'semiotic' approach were deeper; Ginzburg traced them back to forms of explanation and divination that could be oriented towards past, present or future (jurisprudence, medicine and divination proper). Furthermore, his hypothesis was that the origin of the diagnosis from signs or symptoms lay in the practices of long-ago hunters and the 'reading' of animal tracks.

This kind of knowledge based on conjecture and speculation (born of experience, of the concrete and individual) responded to a paradigm that differed from the more prestigious scientific one, but it was used by all kinds of people. In the 18<sup>th</sup> century the situation changed when the bourgeoisie appropriated for itself much of the knowledge of artisans and peasants. The *Encyclopédie* is signalled by Ginzburg as the symbol and chief instrument in this offensive, with the novel and the literature of imagination as a substitute and reformulation of initiation rites, giving access to experience in general. Because of all this, the conjectural paradigm enjoyed an unexpected success. In addition, in the 18<sup>th</sup> and 19<sup>th</sup> centuries the constellation of conjectural disciplines changed, many new ones were born, with medicine assuming a preminent position among them. All the 'human sciences' attempted to relate themselves to it explicitly or implicitly, and they did so by accepting the medical conjectural paradigm of semiotics. Medicine, and thus symptom deciphering, was well known by all the three authors mentioned by Ginzburg as well as by Peirce [33]. This knowledge probably helped them to formulate their methods according to the conjectural paradigm of medicine in a more accurate and convincing way. In so doing, their contributions to their disciplines gained a better 'methodological reputation' so to say.

Many of the controversies related to authorship identification of artworks (the main issue addressed by *connoisseurs* like Morelli) use the two types of hypothetical reasoning referred to by the historian of science Lipton [27]. He distinguishes between inference to the likeliest and to the loveliest explanation.

It is not clear whether the inference about the question of authorship precedes explanation or not. The use of inference to the best explanation (IBE) in the case of authorship identification and, more generally, in the study of visual culture, inverts the usual point of view about the relationship between inference and explanation. According to the natural point of view, or to common sense, inference would precede explanation. In spite of this, the reasoning model implicit in the 'Morelli method' consists of analysing to what extent the evidence can explain a set of hypotheses. In this model therefore, IBE, and thus the explanation, comes before the inference.

Perhaps because of the impact of Ginzburg's essays, the 'Morelli method' is usually the only one mentioned when abductive reasoning is presented in relation with art history. Moreover Morelli is generally the only reference cited to the studies on art when the influence of Peirce on contemporary thought is debated. For further details see Laine Ketner [24]. Besides the influence of structuralism and poststructuralism on the work of many art historians and specialists in the field of cultural and visual studies, from the second half of the 20<sup>th</sup> century on, authors like Holly [23] have noted that some of the issues that were addressed by early semioticians were already being explored at the same time by art historians like Riegl and Panofsky. According to Holly Panofsky was a keen student of semioticians' works and shared certain epistemological predispositions with semiotics. For Argan [4] Panofsky's method, iconology, confronted the problem of art as that of linguistic structures much more than the formalism of Wölfflin. Perhaps that is why Argan affirmed that Panofsky was the Saussure of art history. Although, as Hasenmueller [22] has noted, there are problems in simply calling Panofsky's work semiotic, as semiotics and iconology have a common interest in uncovering the deep structure of cultural products. Iconology, like early semiotics was devoted to exposing the existence of the conscious and unconscious rules of formation that encircle a language and make possible its sudden emergence -both visual and linguistic- on the surface of human history. For further details see Holly [23].

But what interests us here is that Panofsky's writings can be taken as an index of how he reached his conclusions. Panofsky's objective remained the value judgment he called 're-creative synthesis'. For him the definition of an artwork as a 'man-made object demanding to be experienced aesthetically' confronted the researcher with what he considered was the 'basic difference between the humanities and natural science'. The scientist dealt with natural phenomena and could at once proceed to analyse them. In contrast the humanist dealt with human actions and creations and had to engage in a mental process of a synthetic and subjective character. Humanists had 'mentally to re-enact the actions and to re-create the creations', and it was by this process that the real objects of the humanities came into being. According to Panofsky [32] the object of the humanities, and more precisely that of art history, was the result of this re-creative synthesis which was always in process. That is why he explained that the art historian did not constitute

his object through a re-creative synthesis first, followed by archaeological research. For him these two stages did not occur successively, but took place rather in an interwoven manner: the re-creative synthesis served as a basis for the archaeological research, but the latter served in its turn for the process of re-creation. Both stages were only conceived separately in theory (as a way to explain his method) but in practice they were recognised and used to qualify and correct each other in a reciprocal relationship.

This process is analogous to the abductive reasoning model described by the Dutch linguist Gorleé [20] as a method in interlinguistic translation. The necessary application of this method is manifest in the case of descriptive translation, whose objective is translation as a product. As an example, she mentions within this category translation understood as transference. There similarities are recognised that justify a translation which is considered valid in a transitory or derived way because words refer to specific cultural activities. In this sense, the explanatory hypothesis used in previous steps affects further research and interpretation procedures. That is why some authors, such as Tursman [42] consider that the use of abductive processes in this kind of studies is better described with the explanatory metaphor of the figure of the helicoid than with a linear figure. This is because there would be always something to go back to, something that could, in some way, be rediscovered. Gorleé [21] affirms that Peirce's logic-semiotic method can be fully applied to the identification, description and analysis of translation as a mental experiment in the generation of meaning, where a hypothesis generated by abduction is verified in a reiterative way.

The helicoid figure referred to by Gorleé can help us to evaluate the significance of abductive reasoning in the cases of Panofsky and Morelli. On the one hand, the affirmations of the latter are based on an abduction process that goes from effects to possible causes. On the other hand, abduction in Panofsky is linked to belief revision. In other words, it has to do with an understanding of abductive reasoning as ampliative and non-monotonic. It is evident that in both authors the use of abduction led them to infer hypotheses that could not be classically deduced from the given facts. In spite of that, Morelli used abductive reasoning to make irrefutable statements on the authorship of artworks just as if they were the result of deduction. So, even if both Morelli and Panofsky used the same type of reasoning they did not evaluate in the same way the impact of their statements on the discipline of art history. Morelli was deeply fascinated by the power of apodictic demonstrations of objects whose meaning, in fact, is partially veiled as are the objects themselves. In contrast, some of Panofsky's affirmations indicate that he was more aware of the always-in-process nature of interpretations of cultural objects.

The authors Kohlas, Berzati and Haenni [25] affirmed that abductive explanations are in general neither complete nor sound, and that for this reason they are not fully appropriate for model-based diagnosis. Nevertheless model-based diagnosis has been used in combination with abductive reasoning in

many research projects that deal with medical diagnosis. We share to some extent the scepticism of these authors, and propose the potential of the analysis of disciplines that study visual culture to analyse model-based abduction.

The paradoxical situation of objects whose meaning is always partially veiled can be better understood if we turn to computational studies. According to Thagard [41], this field provides a model for a better understanding of the hidden meanings of the data themselves and of the hidden meaning given to them by the producers of those data. Thagard distinguished four types of abduction: simple (which produces hypotheses about individual objects); existential (that postulates the existence of previously unknown objects); rule-forming (that produces rules that explain other rules), and analogical (that uses past cases of hypothesis formation to generate hypotheses similar to existing ones). But it would be difficult and inconsistent to classify the use of abduction in the construction of interpretation of cultural objects (by Morelli, Panofsky or any other interpreter) in just one of these four types.

Abduction is described as a useful mechanism for explaining knowledge acquisition in areas where empirical methods for testing hypotheses are not available, hypothesis, for example, about past or unique events. This inferential process is irreducible to other types of inference as Hintikka affirms. It has been used to describe the cognitive processes that intervene in scientific discoveries in experimental sciences. For further details see Rivadulla [38]. Although the link between this reasoning model and experimental sciences is unquestionable, we think that it has been overvalued. This is evident if we take into account the fact that scientific discovery and the logic of invention are not exclusive of experimental sciences. If abduction is a particular type of argument or epistemic process that attempts to model the incorporation of new beliefs as Aliseda [1] maintains, this process would be one of the principal characters in other kinds of research such as the study of visual culture.

These pages try to explore this tentative hypothesis by presenting analogies between the field of art history and visual studies, and abductive cognition. This is so because topics, inquiries and controversies in these disciplines could not exist independently from the three types of hypothesis identified by Peirce. In any case, they refer to facts or entities unobservable when the hypothesis was formulated but observable later; or to entities or facts that someone could observe in the past even though it is not possible to repeat the observation now, because they are facts of the past; or to entities unobservable in practice. But analysis of studies of visual culture in the light of abductive cognition is not only based on the Peircean definition of the types of hypothesis. Peirce [33] also stated that all thinking is in signs, and signs can be icons, indices, or symbols. All inference is a form of sign activity, where the word sign includes feeling, image, conception, and other representation. Along with these two arguments (one dealing with the different types of hypothesis, and the other with inference as a form of sign activity), a third can be found in Magnani [28] and Magnani and Li. Ping [30]. This author introduces the concept of theoretical and manipulative abduction. He maintains



that there are two kinds of theoretical abduction: sentential, related to logic and to verbal/symbolic inferences, and model-based related to the exploitation of models such as diagrams, pictures, etc. He reminds us that Peirce considered any cognitive activity whatever to be inferential. This included perceptual knowledge and subconscious cognitive activity, not only conscious abstract thought.

### 3 Empathy as Embodied Mechanism

Elements in the style of paintings were considered by Morelli, his heir Bernard Berenson and other *connoisseurs* as unconscious marks that identified their authors. The idea behind the assumptions of these *connoisseurs* was, as Friedländer [13] pointed out, that creative individuality had an unchangeable core and that the artist remained fundamentally the same. Something, therefore, that could not be lost revealed itself in his very expression. In spite of this, just as experience has shown (a well known example of this being the development of the Rembrandt Research Project), this assumption has to be taken carefully, as nothing prevents an artist from switching consciously between different styles in a way similar to the choice of high or low style of a rhetorician, according to the particular occasion of his speech.

In the writings of the scholars known as formalists, style was important not because it was considered characteristic of an individual artist but because it was understood as the specific expression of an age. The most significant representatives of the formalist stream, Riegl and Wölfflin, argued that art offered unmediated sensory access to past world-views. If, according to Ginzburg, Morelli took a prestigious model such as medicine to support his attributions, the formalist authors and their interest in physiology and psychology can be said to respond to a similar aspiration to gain theoretical authority.

The authors and ideas that influenced formalist art historians most were the German physicist and physiologist Hermann von Helmholtz, the psychologists Joahnn Fredrich Herbart, Theodor Lipps and Wilhelm Wundt, the aesthetic theory of the sculptor Adolf Hildebrand and Konrad Fischer. Their views formed the basis of the way Riegl and Wölfflin understood art and its changes over time. They thought that the development of art through history responded to a process of development of vision that was analogous to the development of psychology of perception in individuals. By studying vision and the history of perception these authors focused on the relationship that people had to their environment. For them physical involvement in artworks provoked a sense of imitating the motion seen or implied in the work, and this enhanced the spectator's emotional responses to it. This idea was the result of the influence of empathy theory on the work of these art historians. The fundamental doctrine of empathy theory was that aesthetic experience depended on the experiencing subject's projection of bodily sensations and

emotional memory on fundamental formal elements of experience, such as lines and colours, and thus justified the interest in and need of formalist analysis. Vischer [45] was the first to employ the term ‘Einfühlung’ in a doctoral thesis, meaning the physical responses generated by the observation of paintings. Afterwards, Theodor Lipps, promoted this term and empathy theory in works such as *Die ästhetische Betrachtung und die bildende Kunst* [26]. Lipps was the supervisor of Wölfflin’s dissertation *Prolegomena zu einer Psychologie der Architektur* [51] where the latter gave an ahistorical account of how architectural forms are perceived. Following Lipps’ ideas, Wölfflin stated that forms had no expression by themselves. In reality this only happened when the viewer read the proportions and relations of forms according to his own physiological and psychological constitution, endowing them with something of his own body’s posture and mood.

In spite of the early influence it had on his work, Wölfflin would progressively move away from empathy theory in order to explain stylistic changes through time. In *Renaissance und Barock* [50] he affirmed that changes in style and in other spheres of life as well occurred because of changes in bodily feeling. Later on, in *Classic Art*, he maintained that styles are conditioned by the combination of two independent factors: changes in purely artistic forms of vision, and changes in feelings and states of mind. Finally in his most famous book *Kunstgeschichtliche Grundbegriffe* (Principles of Art History) [49] he proposed a general set of descriptive terms to capture the artistic visual forms of an age without proposing any further explanation. In the introduction he criticised empathy theory arguing that when forms are read as expressions of states of mind, we make the false assumption that the same expressive methods are always available.

Following a process opposite to Wölfflin’s, another important formalist author Riegl rejected the application of empathy theory to art history in *Stilfragen* (Problems of Style) [36]. His later work, however, would show implicitly that he had come closer to it. For example, in *Spätromische Kunstindustrie* (Late Roman Art Industry) [35], where he adopted the distinction between tactile and optical perceptions, he accepted the assumptions of empathy theory when he made the analogy between the apprehension of individual objects in the early haptic stage, and the sense we have of our own bodies. His last major work *Das Holländische Gruppenporträt* (The Group Portraiture of Holland) [37] focused on the paintings’ implicit viewer and this brought Riegl closer to empathy theory. For Riegl Dutch paintings achieved coherence only when the viewer involved himself with the psychic sphere represented in them. According to this author, art in Holland was objective because it was concerned with the psychological relationship between figures that were independent from each other and from the viewer, a relationship that took place at a particular moment in a particular place in the absence of the artist’s subjective point of view.

Empathy was also among the interests of another major figure in the study of the arts, Aby Warburg. He thought it was possible to demonstrate, for

specific conditions of time and place, how the visual arts expressed the perceptions and experiences of man. He analysed the representation of the movement of the body, hair and garments in artworks of 15<sup>th</sup> century Florence and traced back those movements in ancient art and also in contemporary images. For Warburg the borrowing of artistic forms from Antiquity had to do not just with forms, but was justified in terms of an affinity of expressive need. The intensified mimicry of Antiquity, its postures and gestures, were interpreted by Warburg as traces of violent passions experienced in the past, which were used by following generations as a repertoire to represent specific states of action and psychological arousal.

Warburg called these *Pathosformel* ('pathos formula' or 'emotive formula') a name that emphasized the stereotypical and repetitive aspect of the imagined subject the artist had to use to give expression to 'life in movement'. This term appeared for the first time in his essay on *Dürer and Pagan Antiquity* where Warburg [48] traced back the iconographic theme of Dürer's etchings *Orpheus* to the 'pathetic gestural language' of the art of antiquity. He discovered and traced this *Pathosformel* by scrutinizing all relevant evidence: archives, family diaries, psychology, folklore, mythology, religion, philosophy, ethnography, opera, astrology, etc. He even travelled to New Mexico to witness the 'living paganism' of the Pueblo Indians. All these interests gave form to the collection of his library, with the Greek inscription *MNEMOSYNH* (Mnemosyne) above the door. As we will see later, the objects he named after Mnemosyne, the mother of all muses, would play a fundamental role in the development of his thought.

In *The Power of Images* Freedberg [12] described some of the recurrent symptoms of emotional responses to paintings and sculptures throughout history. He intended to draw attention towards the lack of interest that the history of art had taken in doing any research on the subject. In that book Freedberg referred to two kinds of response: direct and indirect, or unmediated and mediated. The first type of response seemed to be automatic and to be predicated on immediate or felt bodily responses, and the second type was mediated by concept, reflection and recollection. The first one can be said to be common to all humans, and the other is influenced by social, cultural and historical conditions. Could mediated response be understood as part of Umberto Eco's description of a hyper coded abduction?

To acknowledge the hermeneutic potential of the relationship between the neuronal bases of response and their historical and cultural inflection, Freedberg [10] has engaged in interdisciplinary work with neuroscientists. The objective of this collaboration is to find physical evidence of how art engages with the body and what the emotional responses that may ensue are. Of course, he signals that the question of the relations between inner and outward movement has a long tradition in the history of art and aesthetics (mentioning the previously cited authors among others), and also the interest in the arts of several neuroscientific works, but his intention is to discover the neuroscientific resolution (or at least refinement) of some of the older

intuitions, hypotheses and theories. His current work, therefore, deals with the neural bases of empathy and the relationship between emotional and motor responses to works of visual arts.

He has collaborated with neuroscientists such as Gallese who coined the term 'embodied simulation' to refer to a common functional mechanism that is the basis of both body awareness and basic forms of social understanding [15]. One of the results of this collaboration is a paper on the neural basis of motion, emotion, empathy and aesthetic experience. For further details see Freedberg and Vittorio Gallese [11]. In addition his work with neuroscientists Battaglia and Lisanby [5] examines the corticomotor networks involved in responses to the sight of particular gestures in artworks.

This collaboration between art historians and neuroscientists has challenged the primacy of cognition in responses to art. They propose a theory of empathetic response to artworks that is not purely introspective, intuitive or metaphysical but has a precise and definable material basis in the brain. They maintain that a crucial element of aesthetic response consists of the activation of embodied mechanisms encompassing the simulation of actions, emotions and corporeal sensation. These mechanisms are mirroring mechanisms and embodied simulation for empathetic responses to images in general, and to works of visual arts in particular. This gives importance not only to context and meaning in art but also looks for a response to works of art that is the same for all humans.

If the studies mentioned above are concerned with artworks only, the analysis of the broad field of visual culture as something that is interwoven with the body is one of the recent incorporations in the interests of many researchers. We can see the emergence of this matter in relation to what Moxey [31] signalled as the introduction of the problem of the 'presence' of the objects of visual culture (of their power as agents) when carrying out research on them. As an example of this, the statements of Belting [6] in *Bild-Anthropologie* can be cited. This author affirms that visual artefacts are embedded in mediums and that neither images nor mediums can be studied separately. This idea of medium is a metaphor for the human body: visual artefacts are inscribed in mediums just as inner images are inscribed in the human body. The medium is thus a figure necessary to the agency of visual objects that are conceived as something more than plain representations.

It can be said, however, that a full theoretical development of concepts such as embodied simulation would be possible if a more complex relationship between visual studies and abduction studies were established. This relationship should be established from a philosophical point of view, and also from that of cognitive sciences, psychology of perception and visual argumentation. To some extent this means that concepts like 'embodied simulation', 'empathetic response' or 'Pathosformel' can be presented as interplay between disciplines and, by extension, that both art history and visual studies are a cognitive niche of interdisciplinary research. The interpretation of visual culture can be analysed as a cognitive process that can be applied to an individual, a collective, a

group or a historical period. The activity of interpretation can be presented both as a process and as the result of a process. In both cases cognitive abduction is constantly present and may appear either as theoretical abduction (related to logic and to symbolic inferences), or as model-based related to the exploitation of models (pictures, photographs, diagrams, collages, etc.) or even as manipulative abduction. Perception is a limited process. This implies the use of this type of reasoning, also understood as epistemic change, for modelling the incorporation of new beliefs. The interpretative process and product, the bodily involvement in visual culture experience and even visual culture itself, can be understood as products that have an inferential structure or that even imply an inferential play. Hence studies in abduction cannot be indifferent to visual studies. The total evidence principle referred to by Eco (that it is impossible to register all the potentially relevant information) transforms perception into an abductive activity in itself. There is evidence for the consistency of this approach. The development of 'image-based hypothesis formation' has led authors like Magnani to consider abduction in terms of visual abduction. But the integration of visual abduction in the study of visual culture invites to explore a path where there is still much to discover.

#### 4 Manipulative Abduction and *Mnemosyne*

To many of the authors who have stressed the agency of visual culture, the figure of Warburg emerges as some sort of 'historiographic hero'. For further details see Moxey [31]. In the field of archaeology, Shelley [39] stressed the important role played by the representation of visual images in the construction of new hypotheses. Abductive reasoning is constantly used in archaeology to discover archaeological remains and archaeological complexes. In the case of this discipline the discovery of some objects can be taken as an index of the existence of others that are absent. Abductive reasoning in archaeology is used to discover new forms or material remains that would be shaped in different ways depending on the associated assumptions. Abduction is thus related to the form of the objects, to their structure, and to the analogical inferences used in each case.

In Warburg's research abduction is not used as a form of induction as described by Reilly [34], neither it is understood as the inverted *modus ponens* described by Anderson [3]. It is seen as the heuristic form studied by Anderson [2]. Warburg's idea of *Pathosformel*, and his project of image argumentation are based on the assumption that the heuristic he proposed helped to obtain explanations with a certain inferential structure. In this sense, problems in interpreting the meaning of images are similar to those in the interpretation of texts, and of interactive discourse: it is impossible to escape the use of inferential structures. For further details see González Navarro [18]. In both cases the distinction between the hidden meanings of the data themselves and the hidden meanings of the producers of those data is a large theoretical

challenge as it was explained by Gabbay and Woods [14]. Warburg faced this challenge. We propose here an interpretation of his project of the atlas *Mnemosyne* according to which he offered a particular answer and a specific expression of the theoretical challenge as we described it before.

The atlas of images *Mnemosyne* was the last ‘tool to see time’, the last device Warburg worked on between 1924 until his death in 1929. It was based on the intuition that a regulated redistribution, a *problematized remontage* of the materials assembled during 30 years of research, could be great, heuristic fertility. This atlas of images was thought in connection not only to the theoretical manuscripts that accompanied the atlas elaboration, but also to the books of Warburg’s library. For Warburg his library was not an ivory tower but an experimental device that made out of the Warburgian *Denkraum* a laboratory where machines to see time could be invented through action on words, images and gestures. The organization of the books in the library was designed by Warburg himself so that the reader would find not only the books she or he was looking for, but also their unexpected ‘good neighbours’. The black panels of the atlas *Mnemosyne* were a place where images were disposed and composed and they constituted crucial elements in Warburg’s talks. He was worried about how to present an argument whose elements were not words or propositions but images that were distant in space and time. As we said before, the atlas was an experimental device, a type of device where the lecturer and his audience were surrounded by a multiplicity of images that acted as visual indicators and not just as illustrations in the exposition of the argument.

Didi-Huberman [8] affirms that Warburg found in the atlas *Mnemosyne* the device that his investigation had always been waiting for: a method capable of manipulating as *interpreting objects* the images that themselves constituted the *objects to be interpreted* in the first instance. The Warburgian analytical space is based on a search for truth that transgresses the frontiers of knowing and seeing, of discourse and image, of the intelligible and the sensitive. But also because of that, it transgresses the canonical and deterministic models of explanation. According to Didi-Huberman, *Mnemosyne* is a theoretical work based on challenging the erudite explanation. It appears as a visual installation where that which cannot be explained in a deterministic way will have to be shown, where an *Übersicht* (a synoptic view) could go beyond univocal propositions, and establish a proper vision of the world. To put it in different words, the atlas *Mnemosyne* was an ‘*übersichtliche Darstellung*’. At the same time that Warburg established his practice, Wittgenstein established his reason, a synoptic presentation of multiplicity valuable because of its heuristic capacity to raise comparisons.

Atlas *Mnemosyne* is characterised by Didi-Huberman as inexhaustible because of its capacity to mount, dismount and remount constantly a corpus of heterogeneous images in order to create unknown configurations and apprehend thanks to them unnoticed affinities or existing conflicts. *Montage* has to be understood as a procedure that goes beyond the artistic practice and

is able to open new spaces of thinking. As a consequence of all this, *montage* reveals itself as a very useful and significant space in epistemic terms. It is useful because it offers the spectator the possibility to conform, acquire or select beliefs, and significant because that space is clearly inclined towards an agent's epistemic stage conceived as an individual activity that models it as a consistent set of beliefs that can change by expansion and contraction.

According to this conception of belief revision the message, or in this case the interpretation, has priority over the agent's initial belief. However the progressive observation of more elements demands the use of an abductive reasoning process that finally turns into an operation that allows the emergence of observations oriented to the epistemic change of our beliefs or interpretations about the objects created by *montage*. In these spaces there exists constantly and for each agent what Aliseda [1] has called abductive novelties (that cause abductive expansion), and abductive anomalies (that can imply the revision of previous beliefs or interpretations). That is the basis of the rational foundations of the heuristic *montages* we are examining and interpreting as if they were situated and embedded cognitions. Because of it, this exhibition space can be understood as an invitation to explore cognition understood as abductive cognition. As Walliser [46] affirms abduction leads to the inference of hypotheses that cannot be classically deduced from the given facts. Objects and spaces constructed by *montage* cannot be interpreted as necessary deductions, they open a space for creativity and therefore, for abduction. Inherent to *montage* is the assumption that abduction is a model of epistemic change. Any individual, group or collective that places itself inside this space will have to interpret through abduction.

The assumptions that encourage this conception of epistemic change conceive of action as a device that provides otherwise unavailable information so that the agent is able to solve problems by performing an abductive process of generation or selection of hypotheses. Because of this, *montage* can be defined as a mechanism that reinforces epistemic change through manipulative abductions. In this type of abduction exemplified by *montage*, inferences are mediated through actions that create external objects which produce new affordances and through the detection of past affordances.

Warburg's atlas *Mnemosyne* is a very clear example of the combination of theoretical and manipulative abduction in the studies of visual culture but we think that, in fact, this kind of reasoning is used continuously in these studies. Could it be affirmed that *montage* is one of the basic activities performed (physically and/or mentally) by specialists in visual culture?

It is unlikely that the disciplines concerned with the study of visual culture can avoid the controversy between the supporters of internal cognition, who think that psychological processes do not extend outside the head and can be explained in isolation from their environment, and those of embedded cognition, according to whom cognition depends on external props and the structure of the environment. For further details see Sprevak [40]. Our objective here was not speculate or to adduce reasons for and against one position or the other. We

have presented analogies in order to show that studies in visual culture have to be seen as a field where cognitive abduction can be explored in the light of a broad epistemic perspective.

Nevertheless there is an assumption in the field of studies dealing with visual culture that has to be stated specifically. The problem of interpretation seems to be deeply rooted in these disciplines. This may be true, but that should also be the place assigned to abduction if we understand it as inseparable from the cognitive process by which we produce and revise interpretations. As a result, abduction could be presented as the logical pattern inherent in interpretation, thus answering one of the unresolved questions of the so called philosophy of interpretation.

The integration of the tradition of the studies in visual culture (represented by art history and visual studies) into studies of abduction would mean the introduction of an interpretative phenomenon that clearly reunites the representational and inferential components present in reasoning. Brandom [7] pointed out the differences between the position of Descartes and Leibniz in Enlightenment. On the one hand, Descartes divided the world into *res cogitans* and *res extensa*, thus converting the possession of representational contents into an explanatory but inexplicable instance. In contrast, Leibniz and Spinoza were concerned with what indicated the fact that a thing represented another taking into account the inferential significance of the representation. This should be elucidated through inferential relations. One of the main challenges since then has been to find how to define representational properties according to inferential ones. Abduction is part of this controversy, and it transforms radically the notion of 'interpretation' as González Navarro has stated [19, 17]. The consideration of the correctness of an inference is not a logical or a formal one; it is a hermeneutic matter, pragmatic and contextual. As Vega Reñón argues [44], the legitimacy of an inference manifests itself in relation to the set of beliefs actualised by the agent in order to cope with a situation. In this sense, the success of an inference depends on the intentional and epistemic attitudes of the agent. As a result, the justification of the inference becomes as complex as the rationalising of human action can be.

The inferential pattern of abduction would harmonize with interpretation understood as a form of cognition that is used, for example, in the production of new interpretations or even in the production of hypotheses leading to the development of new theories. Hence, the production of theories is an intrinsically interpretative process (conceive,  $T$ ; transform  $T$  into  $T_1$ ; extend  $T_1$ ; reject  $T_1$  in favour of  $T_2$ ; producing, then,  $T_n \dots$ ). The acquisition of a language and the historicity of our comprehension preform our cognitions through time individually and collectively. The ampliative effects observed in IBE are the result of the application of a reasoning model that is integrated into the action of interpretation. The inferential parameters that determine the logical relationship between *explanandum*, the *explanans* and abductive explanation are inseparable parts of an abductive competence which is shared in a theoretical and a manipulative sense.



## References

1. Aliseda, A.: *Abductive Reasoning. Logical Investigations into Discovery and Explanation*. Springer, Dordrecht (2006)
2. Anderson, D.R.: *Creativity and the Philosophy of Charles Sanders Peirce*. Clarendon Press, Oxford (1987)
3. Anderson, D.R.: The evolution of peirce's concept of abduction. *Transactions of the Charles S. Peirce Society* 22(2), 145–164 (1986)
4. Argan, G.C.: Ideology and iconology. *Critical Inquiry* 2(2), 297–305 (1975)
5. Battaglia, F., Lisanby, S.H., Freedberg, D.: Corticomotor excitability during observation and imagination of a work of art. *Frontiers in Human Neuroscience* 5, 1–6 (2011)
6. Belting, H.: *Bild-Anthropologie: Entwürfe für eine Bildwissenschaft*. Fink, Munich (2001)
7. Brandom, R.: *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge (1994)
8. Didi-Huberman, G.: *Atlas. ¿Cómo Llevar el Mundo a Cuestas?* Museo Nacional Centro de Arte Reina Sofía, Madrid (2011)
9. Dikovitskaya, M.: *Visual Culture. The Study of the Visual after the Cultural Turn*. MIT Press, Cambridge (2006)
10. Freedberg, D.: Movement, embodiment, emotion. In: Dufrenne, T., Taylor, A.C. (eds.) *Cannibalismes disciplinaires. Quand l'histoire de l'art et l'anthropologie se rencontrent*, pp. 37–61. INHA/Musée du Quai Branly, Paris (2009)
11. Freedberg, D., Gallese, V.: Motion, emotion and empathy in esthetic experience. *TRENDS in Cognitive Sciences* 11(5), 197–203 (2007)
12. Freedberg, D.: *The Power of Images*. University of Chicago Press, Chicago (1991)
13. Friedländer, M.J., Borenius, T.: *On Art and Connoisseurship*. B. Cassirer, London (1942)
14. Gabbay, D.M., Woods, J.: *The Research of Abduction*. Elsevier, Amsterdam (2005)
15. Gallese, V.: Embodied simulation: from neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences* 4, 23–48 (2005)
16. Ginzburg, C.: *Clues, Myths, and the Historical Method*. John Hopkins University Press, Baltimore (1989)
17. González-Navarro, M.: *Interpretar y Argumentar*. CSIC/Plaza y Valdés, Madrid and México (2009)
18. González-Navarro, M.: Intelligent environments and the challenge of inferential processes. *Tijdschrift voor Filosofie* 72(2), 309–326 (2010)
19. González-Navarro, M.: *Hermenéutica*. In: Vega-Reñón, L., Olmos, P. (eds.) *Compendio de Lógica, Argumentación y Retórica*, pp. 271–276. Trotta, Madrid (2011)
20. Gorlé, D.: A eureka procedure: Pragmatic discovery in translation. *European Journal for Semiotic Studies* 8(2/3), 241–269 (1996)
21. Gorlé, D.: *On Translating Signs: Exploring Text and Semio-Translation*. Rodopi, Amsterdam and New York (2004)
22. Hasenmueller, C.: Panofsky, iconography and semiotics. *The Journal of Aesthetics and Art Criticism* 36(3), 289–301 (1978)
23. Holly, M.: *Panofsky and the Foundations of Art History*. Cornell University Press, Ithaca and London (1984)
24. Ketner, K.: *Peirce and Contemporary Thought: Philosophical Inquiries*. Fordham University Press, New York (1995)

25. Kolhas, J., Berzati, D., Haenni, R.: Probabilistic argumentation systems and abduction. *Annals of Mathematics and Artificial Intelligence* 34, 177–195 (2002)
26. Lipps, T.: *Die ästhetische Betrachtung und die bildende Kunst*. Voss, Hamburg and Leipzig (1906)
27. Lipton, P.: *Inference to the best Explanation*. Routledge, London (2004)
28. Magnani, L.: Model-based and manipulative abduction in science. *Foundation of Science* 9(3), 219–247 (2004)
29. Magnani, L. (ed.): *Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Berlin and Heidelberg (2009)
30. Magnani, L., Ping, L. (eds.): *Philosophical Investigations from a Perspective of Cognition*. Guangdong People Publishing House, Guangzhou (2006)
31. Moxey, K.: Visual studies and the iconic turn. *Journal of Visual Culture* 7(2), 131–146 (2006)
32. Panofsky, E. (ed.): *Meaning in the Visual Arts*. University of Chicago Press, Chicago (1983)
33. Peirce, C.: *Collected Papers*. Harvard University Press, Cambridge (1965)
34. Reilly, F.E. (ed.): *Charles Peirce's Theory of Scientific Method*. Fordham University Press, New York (1970)
35. Riegl, A.: *Late Roman Art Industry*. Giorgio Bretschneider Editore, Rome (1985)
36. Riegl, A.: *Problems of Style*. Princeton University Press, Princeton (1992)
37. Riegl, A.: *The Group Portraiture of Holland*. Getty Research Institute for the History of Art and the Humanities, Los Angeles (1999)
38. Rivadulla, A. (ed.): *Éxito, Razón y Cambio en Física. Un Enfoque Instrumental en Teoría de la Ciencia*. Trotta, Madrid (2004)
39. Shelley, C.: Visual abductive reasoning in archaeology. *Philosophy of Science* 53, 278–301 (1996)
40. Sprevak, M.: Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science* 41, 353–362 (2010)
41. Thagard, P.: The best explanation: Criteria for theory choice. *The Journal of Philosophy* 75, 76–92 (1978)
42. Tursman, R. (ed.): *Peirce's Theory of Scientific Discovery*. Indiana University Press, Bloomington (1987)
43. Vega, J.: Del pasado al futuro de la historia del arte en la universidad española. *Ars Longa* 16, 205–219 (2007)
44. Vega-Reñón, L.: Inferencia, argumentación y lógica. *Contextos* III(6), 47–72 (1985)
45. Vischer, R.: *Über das optische Formgefühl: ein Beitrag zur Aesthetik*. Ph.D. thesis, University of Tübingen (1872)
46. Walliser, B.: Abductive logics in a belief revision framework. *Language and Information* 14, 87–117 (2005)
47. Walton, D. (ed.): *Character Evidence. An Abductive Theory*. Springer, Dordrecht (2006)
48. Warburg, A.: Dürer and italian antiquity. In: *The Renewal of Pagan Antiquity*, pp. 553–731. Getty Research Institute for the History of Art and the Humanities, Los Angeles (1999)
49. Wölfflin, H.: *Principles of Art History*. Dover Publications, New York (1950)
50. Wölfflin, H.: *Renaissance and Baroque*. Cornell University Press, Ithaca (1966)
51. Wölfflin, H.: *Prolegomena to a Psychology of Architecture*. MIT, Cambridge (1976)

# Understanding Scientific Inference in the Natural Sciences Based on Abductive Inference Strategies

Jun-Young Oh

**Abstract.** The purpose of this study is to understand scientific inference in the natural sciences through the use of abductive inference. Abductive inference enables scientific discovery through creative inference during problem solving. We present the following two research problems: (1) the validity of a scientific inference procedure building on Magnani's research (2001) that employs various strategies and the criterion of hypotheses choice in order to increase plausibility: puzzling observation, abduction, retrodution, updating, deduction, induction, and recycle; and (2) the validity of our suggested multistage inference procedure for analyzing the "The Return of Halley's Comet" case, which has been called Newtonianism's most public triumph. Through an analysis of a case in the history of science, we describe the patterns of inference and the generation, through available data, of plausible hypotheses based on abductive inference. We then test these hypotheses with the deduction-induction cycle to determine which hypothesis is most plausible. A framework that includes the history of science can potentially provide a more consistent view of scientific practice and promote a deeper understanding of scientific concepts.

**Keywords:** abductive inference, Return of Halley's Comet, history of science.

## 1 Introduction

If a knowledge base does not have all of the necessary clauses for reasoning, ordinary hypothetical reasoning systems cannot explain observations. In this case, it is necessary to explain such observations through abductive reasoning, supplemental reasoning, or approximate reasoning (Abe, 1998).

---

Jun-Young Oh

Hanyang University, Seoul, 133-791, Republic of Korea

e-mail: [Jyoh3324@hanyang.ac.kr](mailto:Jyoh3324@hanyang.ac.kr)

In addition to deduction and induction, Charles S. Peirce argues for a third mode of inference, which he calls “hypothesis” or “abduction”. He characterizes abduction as reasoning “from effect to cause” and “as the operation of adopting an explanatory hypothesis” (Niiniluoto, 1999a). Abduction is more frequently used in everyday “common-sense” reasoning and in expert-level problem solving than is generally recognized (Peng and Reggia, 1990, p.2). Abductive inference additionally enables scientific discovery through creative inference during problem solving (Martin, 2007).

Thus, the primary aims of this study are to develop a program that uses scientific inference processes based on abduction, “the procedure of forming an explanatory hypothesis”, and to explore how abductive strategies have been used in the history of science. To accomplish this objective, we present the following two research problems: (1) to develop a scientific inference process for “the procedure of forming an explanatory hypothesis”; and (2) to apply our suggested multistage inference procedure to the discovery of “The Return of Halley’s Comet”, called the most public triumph of Newtonianism.

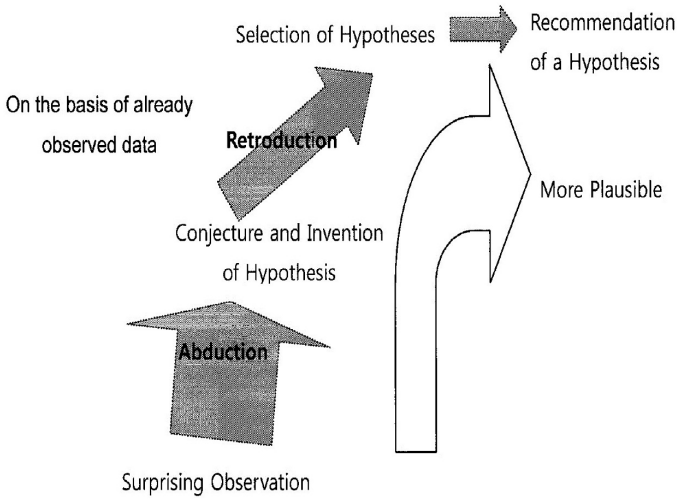
## 2 Theoretical Background

### 2.1 *Abduction*

In Peirce’s 1903 lectures on pragmatism, Peirce tentatively presented a pattern for abduction (CP 5.145):

*The surprising fact, C, is observed;  
But if A were true, C would be a matter of course,  
Hence, there is reason to suspect that A is true.*

In the form of inference he describes, C is a statement or set of statements describing some facts, and A is another statement that supposedly explains C. In premise one, two claims are that C is true in the actual world and that C is surprising. The latter claim can be modeled in many ways, one of the simplest being the requirement that C does not follow from our other knowledge about the world. In premise two, Peirce calls A an explanation of C, or an ‘explanatory hypothesis’ (Flash, and Kakas, 2000, p.7). According to Peirce, abduction is “the process of forming an explanatory hypothesis” (5.172) that “must cover all the operations by which theories and conceptions are engendered” (5.590), including the invention of hypotheses and the selection among them of those to consider further (Kapitan, 1997). Thus, we propose that premise one is what is called (a) “Surprising observations”, premise two is (b) “Conjecture and Invention of hypotheses”, and the concluding premise is (c) “Selection of hypotheses”.



**Fig. 1** Abduction-retroduction cycle

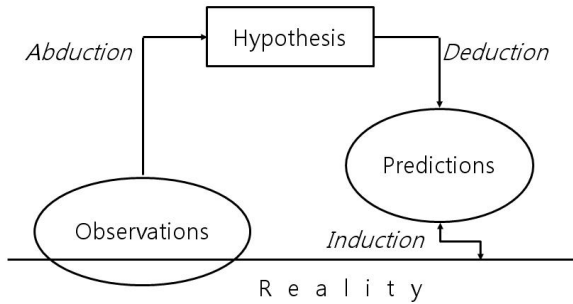
According to Rescher (1978, p.8), Peirce saw qualitative induction as an evolutionary process of variation and selection. The following two component processes have been observed:

- (i) Hypotheses-projection or abduction: the purely conjectural proliferation of a whole gamut of relatively plausible alternative explanatory hypotheses.
- (ii) Hypothesis-testing or retroduction: the elimination of hypotheses on the basis of observational data.

The overall process results in scientific inquiry that repeatedly eliminates rival hypotheses to select one preferred candidate. Each stage of the abduction-retroduction cycle further reduces a cluster of conjectural hypotheses to an accepted theory.

## ***2.2 The Epistemological Model of Hypothetical Reasoning Involving Abduction and Induction***

Abduction is a first phase of inquiry with which ideas are generated. Differing from the evidential viewpoint, however, the methodological viewpoint emphasizes that abduction is one phase in the process of inquiry; hypotheses and ideas are generated with abduction and should then be tested with deduction and induction (Chiasson, 2005). In Peirce's theory of reasoning, Peirce abandoned the idea of a syllogistic classification of reasoning. Instead, he identified the three reasoning forms- abduction, deduction and induction- with the three stages of scientific inquiry: hypothesis generation, predictions, and evaluation as shown in Figure 2 (Flash, and Kakas, 2000, p.7).

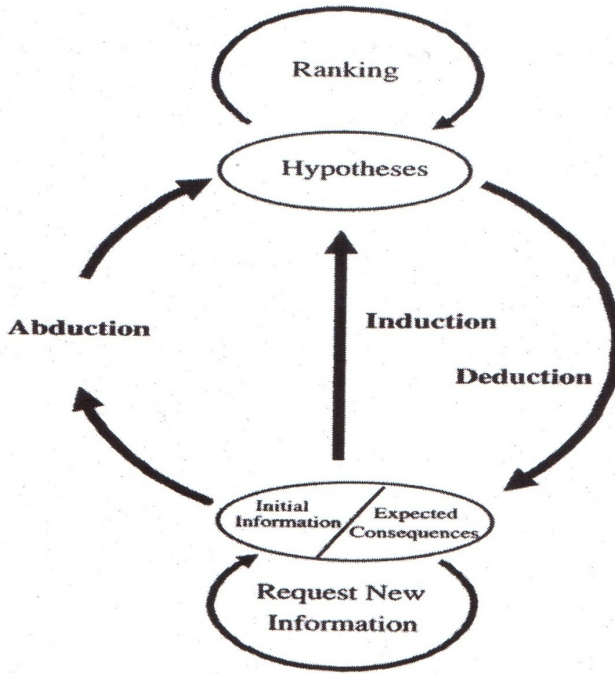


**Fig. 2** The three Stages of Scientific Inquiry (Flash, and Kakas, 2000, p.7)

Magnani and others (Magnani, 1992; Ramoni et al., 1992) developed an epistemological model of medical reasoning called the Select and Test Model (ST-Model) that parallels the classical model of abduction, deduction and induction: the ST-Model describes the different roles played by these basic inference types in various kinds of medical reasoning (diagnosis, therapy planning, monitoring). The ST-Model can be extended, however, and analyzed as an example of scientific theory change.

A hundred years ago, Peirce interpreted abduction essentially as an inferential creative process for generating a new hypothesis. The two main epistemological meanings of the word “abduction” are the following (Magnani, 1988): 1) abduction that generates plausible hypotheses (selective or creative) and 2) abduction that infers the best explanation by evaluating hypotheses. All we can expect of “selective” abduction are hypotheses for further examination; these hypotheses have some chance of turning out to be the best explanation. Selective abduction will always produce hypotheses that are at least partially explanatory and therefore have “a small amount of initial plausibility”. The syllogistic view advocated by Peirce, in which abduction is inference to the best explanation, requires that the final chosen explanation be the most plausible. “In the latter sense the classical meaning of selective abduction as inference to the best explanation is described in his epistemological model by the complete abduction-deduction-induction cycle”. (Magnani, 1999, pp.220-222).

Thus, selective abduction is making a preliminary guess to produce a set of plausible diagnostic hypotheses, and the preliminary guess is then followed by deduction of the consequences of the hypotheses and induction to test them with available data. This process is meant to achieve one of the following objectives: (1) to determine the likelihood of a hypothesis by noting the evidence that it explains better than competing hypotheses, or (2) to refute all but one hypothesis.



**Fig. 3** Epistemological model of hypothetical reasoning (Magnani, 2001. p.74)

If, during this first cycle, new information emerges, hypotheses not previously considered can be suggested, and a new cycle occurs, as shown in Figure 3.

Abe, A. (1998) proposed a combination of deduction and induction to generate humorous conversation where the structure of a previous conversation (deduction) is referenced to generate (abduce) a new conversation. Actually, new observations are obtained as a result of deduction, but his proposal is to generate new observations to conduct proper abduction. That is, deduction is performed when proper observations cannot easily be obtained for abduction.

## ***2.3 The Criteria for Hypotheses Selection***

### **2.3.1 The Criteria for Hypotheses Selection**

Explanatory criteria are needed because rejecting a hypothesis requires that a competing hypothesis provide a better explanation. Clearly, conclusions are reached according to rational criteria, such as consilience or simplicity in some cases, as when choosing scientific hypotheses or theories where the role of “explanation” is dominant (Magnani, 2001, p.27). Consequently, to achieve

the best explanation, it is necessary to have or to establish a set of criteria for evaluating the competing explanatory hypotheses reached by creative or selective abduction (Magnani, 2001, p.26).

Thagard (1978) discusses a static notion of the consilience” of theories, which presupposes that all classes of facts, the total evidence, are given. This is generally how it appears when a scientist presents the results of his/her research. Arguments supporting the superiority of an explanation depend on a range of facts.

Dynamic consilience can be defined in terms of consilience: theory T is dynamically consilient at time n if at n it is more consilient than it was when first proposed, that is, if there are new classes of facts that it has been shown to explain. It is difficult to precisely state a comparative notion of dynamic consilience. Roughly, T1 is more dynamically consilient than T2 if and only if T1 has succeeded in adding more to its set of classes of facts explained than T2 has. Successful prediction can often be understood as an indication of dynamic consilience, provided that the prediction concerns matters that the theory used to make the prediction had not previously dealt with and that the prediction is also an explanation. Successful prediction in a familiar domain contributes relatively little to the explanatory value or acceptability of a theory. In the conservative understanding of dynamic consilience just described, no modification to the theory T or set of auxiliary hypotheses A is needed to explain the new phenomenon. However, a theory will often impress by explaining, through a change in T or A, a phenomenon inexplicable by the previous theory. Thagard used the term radical dynamic consilience to describe this property of theories that succeeds in explaining new kinds of facts by means of modifications of the theory or auxiliary hypotheses.

Accordingly, we must require that the modified theory prove to be conservatively dynamically consilient. The hypothetico-deductive method neglects this dynamic feature of theory evaluation.

To this point, Thagard (1978) treats consilience as a property of theories, but generalizations can also be inferred as the best explanations. One final remark on consilience would be that it appears that a maximally consilient hypothesis or theory explains any fact whatsoever. This could be achieved by a sufficiently flexible set of auxiliary hypotheses to ensure that any phenomenon could be explained by the theory. 'Simplicity' deals with the problem of the level of conceptual complexity of hypotheses with equal consilience. This evaluation is strongly influenced by Ockham's razor: simplicity can be highly relevant when discriminating between competing explanatory hypotheses (Magnani, 2001, p.26). Peirce introduced a principle of "economy" that includes the application of Ockham's razor: "Try the theory of fewest elements first; and only complicate it as such complication proves indispensable for the ascertainment of truth" (1960, 4.35).



### 2.3.2 Non-monotonic Inference for a New Cycle:

In classical logic, a system increases its stock of truths as knowledge is added and as inferences are made. There is no mechanism for discarding information or revising beliefs. This aspect of classical logic is termed “monotonic”. In non-monotonic systems, inferences can be made on the basis of available data, but these inferences can be rejected and new ones made when new data become available (Fischer and Firschein, 1987, p.96).

If new information suggests hypotheses not previously considered, a new cycle of evaluation begins. The cyclical nature of the epistemological model stresses its non-monotonic character. For example, new information can significantly reduce the likelihood of or even invalidate a previous hypothesis (Peng and Reggia, 1990, p.125). Non-monotonic inference is thus time-dependent logic (Trigg, 1991, p.5). Its conclusions must be flexibly revised or retracted when a previously proposed conclusion is contradicted by new information because it makes its conclusion based on typicality or an absence of information about atypicality (Fischer and Firschein, 1987, p.96).

## 3 Scientific Inference Procedure Based on Abductive Inference Strategies Involving the Deduction-Induction Cycle

We suggest a scientific inference procedure building on Magnani’s research (2001) with various strategies and the criterion of hypotheses choice: puzzling observation, abduction, retroduction, updating, deduction, induction, and recycle. We present observations about the use of Halley’s Comet as an example to corroborate Newtonian mechanics.

### 3.1 *Generating Creative Hypotheses*

Stage (1) “Puzzling or surprising observation”.

According to Paavola (2004), strategies are also involved when it is said that abductive inference starts from anomalous or somewhat surprising phenomena. Why is it so often emphasized that abductive inference starts from surprising phenomena (Hoffmann, 1999)

Abductive inference starts from relatively little data, the astonishment phenomenon, to initiate reconstruction strategies (or abstraction; refer to Magnani, 2001, p.72) that differentiate necessary and important data from useless data according to the kind of scientific knowledge available and the features of the problem producing the astonishment phenomenon. This then enables the creation of new hypotheses.

For example, the following occurred with Halley's Comet observations (Giere, 1997):

**(Astonished phenomenon):** Comet's initial observational data. Halley began investigating a comet that he had observed in 1682. These comets were very interesting objects because they had always been viewed as mysterious, even ominous. Their appearances certainly exhibited no apparent regularity (p. 67).

**(Data reconstruction):** If the behavior of a comet were to exhibit an underlying regularity, the comet should have traveled a similar path before, thus eliminating other possible comet movements.

Stage (2) The "Invention of hypotheses" occurs when multiple hypotheses are generated by "analogical abduction strategies". Based on our prior store of declared knowledge in other domains, we used analogical abduction to invent a hypothesis (a tentative explanation) based on existing knowledge in other domains for the puzzling or surprising.

(Invention of hypotheses based on declared knowledge in other domains) Through his observations in 1682, he was probably building on Newton's suggestion that comets may be like small planets with very large elliptical orbits. Indeed, it was impossible to determine from those observations whether the orbit was an ellipse, as Newton suggested, or a parabola (Giere, 1997, p.67).

(Newton's suggestion that comets may be like small planets was more statically consistent with other domains than the other mysterious and even ominous explanations used by the same domains)

Stage (3) "Selection of hypothesis" includes all phenomena present. The first preliminary test stage occurs when tests are planned to select or eliminate hypotheses using retroductive strategies, which are a weak test of a hypothesis because they only determine whether the hypothesis explains the puzzling observation that led to its generation from what we already know in the first place (Lawson 2010). Thargard (1978) discusses a static notion of the consilience of theories, which presupposes that all classes of facts, the total evidence, are given. This is generally how it appears when a scientist presents the results of his/her research. Arguments supporting the superiority of an explanation depend on a range of facts explained.

Selection of an elliptical orbit hypothesis: (retroduction).

Newton's theory allowed the possibility of a parabolic orbit, but such an orbit would mean that the comet passes by only once and then leaves the solar system forever. If, however, the orbit was elliptical, the comet would have traveled that same path many times before (Giere, 1997, p.67).

Halley began digging into the records of observation of previous comets. He found 24 recorded observations, going back roughly 150 years, for which the records were precise enough to compare with the observations of 1682. For two of these, one in 1606-1607 and one in 1530-1531, the recorded orbits were very close to that of the 1682 comet. Halley argued that it was extremely unlikely that three different comets should have such similar orbits and concluded that these were three appearances of the same comet in an elliptical orbit lasting roughly 76 years (Giere, 1997, p.67).

An elliptical orbit hypothesis was more statically consilient than a parabolic orbit

Stage (4) "Updating of the hypothesis" occurs when new hypotheses are generated based on newly available information. The "hypotheses updating phase" is necessary for updating existing hypotheses or generating new hypotheses based on newly available information. This occurs via rule-forming abduction strategies (Thagard 1988, p.5), which consist of focusing on single or paired treatments on the list to perform a more thorough evaluation of their appropriateness to the data at hand. Thagard (1978) treats consilience as a property of theories, but generalizations can also be inferred as best explanations.

According to Peng and Reggia (1990, p.6), it can be concluded based on many studies that human diagnostic reasoning often involves "hypothesis generation" (forming candidate explanations), "hypotheses updating" (updating existing hypotheses based on newly available information), and "hypotheses testing" (disambiguating existing hypotheses). It describes the different roles played by such basic inference types in developing various kinds of medical reasoning (diagnosis, therapy planning, monitoring) but can be extrapolated to illustrate scientific theory change (Magnani, 1999, p19).

Updating existing new hypotheses based on newly available information: He speculated but could not prove that the slight discrepancies in the three orbits were due to gravitational influences from the planets, particularly Jupiter. Halley did not stop there. Using the data from all three cases, together with the hypothesis that he was dealing with a system represented by a Newtonian model .

(Halley's hypothesis, supported by existing knowledge, was more statically consilient than the hypotheses of the previous 76 years)

### 3.2 *Hypotheses Testing: Corroboration of Selective Hypotheses in the Deduction-induction Phase*

The deduction-induction phase involves the actual process of hypotheses evaluation.

A dynamic notion of consilience must also be taken into account when considering the acceptability of explanatory hypotheses. Explanatory criteria are needed because the rejection of a hypothesis requires that a competing hypothesis provide a better explanation. Clearly, conclusions are reached according to rational criteria such as consilience or simplicity in some cases, as when choosing scientific hypotheses or theories where the role of “explanation” is dominant (Magnani, 2001, p.27).

Stage (5) Deduction is connected to prediction. Once a hypothesis for a phenomenon is established, certain predictions derived at time t1 can be revised at time t2.

Scientific research involves raising causal questions about unexplained observations, using abduction to create alternative explanations (alternative hypotheses) and imagining experimental or observational conditions that would allow the deduction of expected outcomes (predictions: expected result)

Stage (6)Induction, which does not mean here an amplitude process of the generalization of knowledge, corroborates those hypotheses whose expected consequences turn out to be consistent with the observed data and refutes those that fail this test. Induction is the final test of an abducted hypothesis; it produces the best explanation by completing the whole cycle of the epistemological model. A new cycle starts if new information suggests hypotheses not previously considered (Magnani 2001, p.73-74).

Andgathering actual outcomes (data: observed results)to compare with expected outcomes, drawing conclusions about the relative support or lack of support for the initial hypotheses based on the quality of the observations and their correspondence with the predictions, and finally, storing supported hypotheses conclusions (Lawson, 1995).

(Deduction): Halley calculated the time of the next return. He boldly predicted that the comet would be seen in late December, 1758. (Expected data).

Observation data: The comet reappeared, as predicted, near Christmas of 1758.

(Induction): The only alternative hypothesis was that another comet with the same orbit just happened to appear right around the predicted time 76 years later. That seemed to everyone extremely unlikely. So, the data provided very good evidence that the Newtonian model fit (Giere, 1997, p.68). (Halley’s hypothesis corroborated by predicted evidence was more conservatively dynamically consilient than it was before he boldly predicted that the comet would be seen).

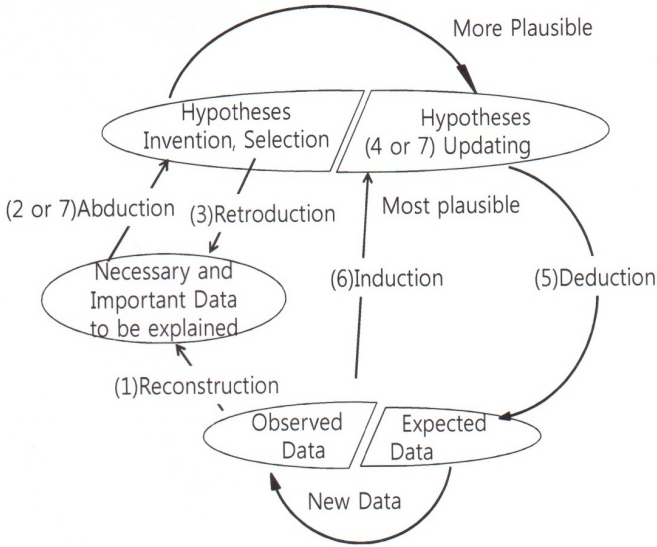


Fig. 4 Scientific Inference Method based on Abductive Inferences

### 3.3 Recycling

#### Stage (7) Abduction (or Updating)-Deduction-Induction Cycle

Upon comparing the prediction with the results obtained, the experimenter may find that the hypothesis has been confirmed, that some necessary modification is indicated, or that it needs to be abandoned.

(Route A): an updating-deduction-induction cycle for no modification to the theory T or set of auxiliary hypotheses A is needed to explain the predictive phenomena for more plausibility of conservative dynamic consilience. Rather than data presentation, it should be predicted deductively by advanced hypotheses that update preliminary test stage hypotheses to form more plausible hypotheses; these new hypotheses lead to predictions deductively through a cyclical updating of hypothetical deduction.

(Route B): a change in T or A is necessary to explain a phenomenon inexplicable by the theory in its original form. It succeeds in explaining new kinds of facts through radical dynamic consilience, revision or retreat in T to explain new data for nonmonotonous inference by a new Abduction-Deduction-Induction cycle.

### **Route A: Updating-Deduction-Induction cycle**

Continuously expansive Newtonian models were applied to fluid mechanisms and other domains. More conservatively dynamic consilience or static consilience than the Newtonian model corroborates Halley's Comet prediction at previous cycle)

Astronomers have long known that the major axis of Mercury's orbit does not remain fixed in space in relation to the stars. The major axis rotates around in the plane of the orbit. Part of this shifting arises from the gravitational attraction of the other planets. When this and other effects are taken into account, there nonetheless remains a residual shift of 41 arcsec per century.

What causes the perihelion advance of Mercury's orbit? Is it perhaps an undiscovered planet, sometimes called Vulcan, orbiting within Mercury's orbit (Newtonian model's auxiliary hypotheses  
No such planet has ever been definitively observed observation unexpected by Newton's theory.

### **Route B: Abduction-Deduction-Induction cycle (non-monotonic cycle)**

General relativity predicts a motion due to the strong curvature of space-time close to the sun (a new hypothesis, general relativity, after retreating or revising Newton's theory). The predicted value for Mercury is 43 arcsec per century, so the observed and predicted results agree to within a few percent. Again, observations confirm general relativity (Zeilik, 2002, pp.141-142). (General relativity is More static than the Newtonian model, corroborated at previous cycle.).

Magnani (2001) defined a "selective abduction" as the process of finding the right explanatory hypothesis from a given set of possible explanations. In this case, we should find the most appropriate rule to construct the conclusion from among the set of rules he has access to. However, it can happen that there is no general rule known to the arguer that would imply the given case. Thus, the arguer must invent a new rule. Eco (1983) calls an abduction that involves the invention of a new rule a "creative abduction". Physicists' attempts to account for the anomalies in the orbit of Mercury provide examples of both undercoded and creative abductions. It is possible to account for the anomalies by making use of the rules already available concerning the motion of planets. One proposed hypothesis of this kind was the existence of an unknown planet close to the sun that was perturbing Mercury's orbit. In this case, the argument is an undercoded abduction. Many such hypotheses were proposed, but in the end, it was a creative abduction, the creation of a new general rule (Einstein's theory of relativity), that successfully accounted for the anomalies (Pedemonte and Reid, 2011).

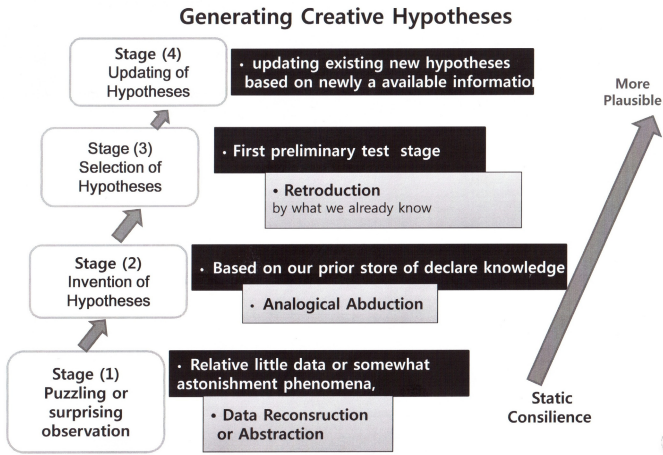


Fig. 5 Generating Creative Hypotheses

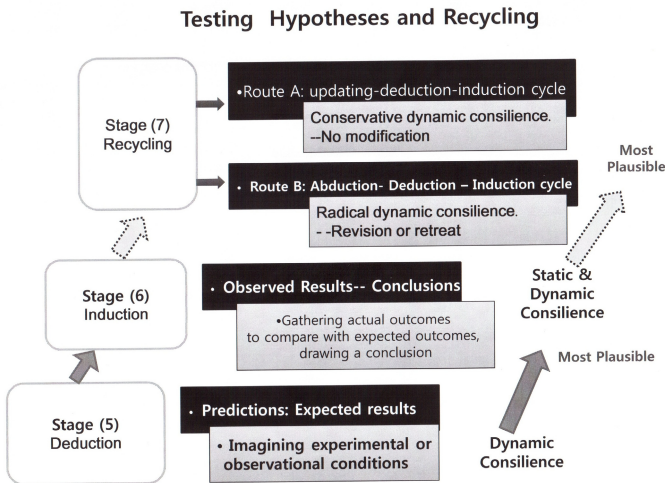


Fig. 6 Testing and Recycling Hypotheses

## 4 Conclusions

We have explored and proposed a scientific inference procedure based on various kinds of abductive strategies for theory choice and examined its validity by applying it to the prediction of the return of Halley’s Comet, upon which Newtonianism won its most public triumph.

First, we suggested a scientific inference procedure based on abductive inference strategies to generate hypotheses involving the deduction-induction method for the evaluation of hypotheses. The term “abduction” is usually applied to the evaluation of explanatory hypotheses, although it sometimes also includes processes of generating them (Charniak and McDermott, 1985; Josephson and Josephson, 1994). The processes, except for that of updating hypotheses, are suggested by this research, as discussed above. But we have revised and enhanced the Magnani research (2001), in addition to offering various strategies and the criterion of theory choice suggested by Thagard (1978, 1988).

We can identify the following pattern of scientific inference and arguments, as depicted in Figure 3:

- (1) Reconstruction (abstraction) can be considered a process of structuring incoming observed data in a small set of necessary and important entities according to the kind of knowledge available in order to abstract puzzling, surprising, and previously unexperienced phenomena in need of explaining.
- (2) Abduction that is based on the analogical strategy involves conjecturing (guessing) and inventing a set of hypotheses originating from puzzling and previously unexperienced phenomena through reconstruction strategies. The hypotheses generated based on analogical abduction are more plausible (statically consilient) than competing hypotheses based only on simple abduction.
- (3) Once hypotheses have been invented, they need to be ranked according to level of consilience (Thagard 1988), which measures how much a hypothesis can explain. Identifying the highest ranked hypothesis can help in planning the evaluation phase, which begins with tests of the preferred hypothesis. According to Rescher (1978), the hypotheses are then tested according to the familiar process of exploiting them as a basis for predictions, which are then checked. Peirce named this process of eliminating hypotheses by experiential testing (p.3) for more plausible (static consilient) hypotheses retrodution. When their levels of consilience are equal, the level of simplicity (Magnani, 2001, p.26) measures the hypotheses’ level of conceptual complexity for the more plausible (simple) of the competing hypotheses.
- (4) Updating the hypotheses involves updating existing hypotheses or generating new hypotheses based on newly available information (Peng and Reggia 1990, p6) to produce more plausible (statically consilient) hypotheses (refer to Figure 5).
- (5) We then use deduction to generate further predictions, which also requires connections in declarative knowledge (Lawson, 2010). The final chosen explanation is the most plausible one “as inference to the best explanation is described in his epistemological model by the complete abduction-deduction-induction cycle” (Magnani, 1999, pp.220-222).
- (6) Subsequently, we make the necessary observations, which matched our predictions (Lawson, 2010). Then, induction is used as the process of reducing the uncertainty of established hypotheses by comparing their consequences



with observed facts (Magnani, 1999, p.221) for more plausible (dynamic consilient) hypotheses

- (7) We might require that the final chosen explanation be the most plausible complete abduction-deduction-induction cycle. Induction corroborates those hypotheses whose expected consequences turn out to be consistent with observation data, and updating-deduction-induction cycles begin to determine more plausible (dynamic and static consilient) hypotheses (Route A: case of observation outcome corresponding with expected result). .

But if, during the previous cycle, new information emerges (Route B), a new cycle (Abduction-Deduction-Induction) begins to determine more plausible (radical dynamic and static consilient) hypotheses for non-monotonic inference. (refer to Figure 6).

Second, the role of creativity in the invention of hypotheses is very important because hypotheses invented by analogical abduction based on puzzling phenomena in other domains are tentative hypotheses for argumentative claims. If new information suggests hypotheses not previously considered, however, a new cycle begins by revising the existing hypotheses. This process is of a “non-monotonic” character.

Third, we understand the patterns for generating more plausible hypotheses through available data based on abductive inference and the process for testing the plausibility of these hypotheses using the deduction-induction cycle. The key distinction between defensible and indefensible inference is that of monotonicity; defensible conclusions may need to be revised or retracted when additional information becomes available.

Finally, the history of astronomy, as the origin of the natural sciences, is subject to methods of inquiry that drive causal explanations based on the historical evidence of natural phenomena.

Hintikka (1999) maintains that regarding the theory of logic and reasoning, especially at the level of introductory textbooks and courses, the study of excellence of introductory textbooks and courses, the study of excellence in reasoning is often forgotten, and the emphasis is on the avoidance of mistakes in reasoning. According to Hintikka, students are not taught how to reason well but are instead only taught to maintain their logical virtue (to avoid logical fallacies and to learn what is and what is not admissible and valid). The focus has been on definitory rules, and strategic rules have largely been neglected. No one is good at logic and reasoning based on knowing only the definitory rules of logic; one must also master the strategic rules.

Therefore, it seems reasonable to conclude that the use of not only definitory rules but also strategic rules is effective in understanding a natural science using the history of science.

**Acknowledgements.** I sincerely thank Professor Sang Wook Yi of Hanyang University (Korea) for encouraging me to write up my work. I would also like to thank Dr. YongHo Kim of KISTI (Korea) and Professor Yonggi Kim of Chungbuk National University (Korea), for their efforts in improving the readability of the manuscript.

## References

1. Abe, A.: Applications of Abduction. In: Proc. of ECAI 1998 Workshop on Abduction and Induction in AI, pp. 12–19 (1998)
2. Abe, A.: Abductive Analogical Reasoning. *Systems and Computers in Japan* 31(1), 11–19 (2000)
3. Charniak, E., McDermott, D.: Introduction to artificial intelligence. Addison-Wesley, Reading (1985); Chiasson, P.: Abduction as an aspect of retroduction. *Semiotica* 153(1/4), 223–242 (2005)
4. Eco, U.: Horns, hooves, insteps: Some hypotheses on three types of abduction. In: Eco, U., Sebeok, T. (eds.) *The Sign of Three: Dupin, Holmes, Peirce*, pp. 198–220. Indiana University Press, Bloomington (1983)
5. Fischler, M.A., Firschein, O.: *Intelligence: the eye, the brain, and the computer*. Addison-Wesley Publishing Company, Inc. (1987)
6. Flash, P.A., Kakas, A.C.: Abductive and Inductive Reasoning: Background and Issues. In: Flash, P., Kakas, A. (eds.) *Abduction and Induction: Essays on their Relation and Integration*, pp. 1–27. Kluwer Academic Publishers, Dordrecht (2000)
7. Giere, R.N.: *Understanding Scientific Reasoning*. In: Ronald, N. (ed.) *9Harcourt Brace College Publishers* (1997)
8. Hintikka, J.: Is Logic the Key to All Good Reasoning? In: Hintikka, J. (ed.) *Inquiry as Inquiry: A Logic of Scientific Discovery*, Jaakko Hintikka Selected Papers, vol. 5. Kluwer Academic Publishers, Dordrecht (1999)
9. Hoffmann, M.: Problems with Peirce's Concept of Abduction. *Foundations of Science* 4, 271–305 (1999)
10. Josephson, J.R., Josephson, S.G.: *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, Cambridge (1996)
11. Kapitan, T.: Peirce and Structure of Abductive Inference. In: Hauser, et al. (eds.) *Studies in the Logic of Charles Sanders Peirce*, pp. 477–496. Indiana University Press, Bloomington and Indianapolis (1997)
12. Lawson, A.E.: Basic Inference of Scientific Reasoning, Argumentation, and Discovery. *Science Education* 94, 336–364 (2010)
13. Magnani, L.: Abductive reasoning: Philosophical and Educational perspectives in medicine. In: Evans, D.A., Patel, V.L. (eds.) *Advanced Models of Cognition for the Medical Training and Practices*, pp. 21–41. Springer, Berlin (1992)
14. Magnani, L.: *Epistemologie de l'invention scientifique*. *Communication and Cognition* 21, 273–291 (1988)
15. Magnani, L.: Model-based Creative Abduction. In: Magnani, L., Nersessian, N.J., Thagard, P. (eds.) *Model-Based Reasoning in Scientific Discovery*, pp. 219–238. Kluwer Academic/ Plenum Publishers, New York (1999)
16. Magnani, L.: *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic/Plenum Publishers, New York (2001)

17. Martin, R.: *The Opposable Mind: Harnessing the power of integrative thinking*. Harvard Business School Press, Boston (2007)
18. Niiniluoto, I.: *Defending Abduction*. *Philosophy of Science* 66(Proceedings), S436–S451 (1999)
19. Paavola, S.: *Abduction as a logic and Methodology of discovery: The importance of strategies*. *Foundation of Science* 9, 267–283 (2004)
20. Pedemonte, B., Reid, D.: *The role of abduction in proving processes*. *Educational Studies in Mathematics* 76, 281–303 (2011)
21. Peirce, C.S.: (CP): *Collected Papers of Charles Peirce, C.S.(CP), Collected Papers of Charles Sanders Peirce*. In: Hartshorne, C., Weiss, P. (eds.) vols. 1–6; Burks, A.W. (ed.) vols. 7–8. Harvard University Press, Cambridge (1931–1958)
22. Peirce, C.S.: *Phänomen und Logik der Zeichen*. Hrsg. und übersetzt von Helmut Paper, Frankfurt/Main, Suhrkamp (1983)
23. Peirce, C.S.: (HP): *Historical Perspectives on Peirce Logic of Science*, Eisele, C. (ed.). *A History of Science*, vols. 2. Mouton Publishers, Berlin (1985)
24. Peirce, C.S.: (EP2): *The Essential Peirce. Selected Philosophical Writings, the Peirce Edition Project* (ed.), vol. 2 (1893–1913). Indiana University Press, Bloomington and Indianapolis (1992–1998)
25. Peirce, C.S.: *Collected papers*. Harvard University Press, Cambridge (1960)
26. Peng, Y., Reggia, J.A.: *Abductive Inference Models for Diagnostic Problem-Solving*. Springer, New York (1990)
27. Rescher, N.: *Peirce's Philosophy of Science*. University of Notre Dame Press, Notre Dame (1978)
28. Ramoni, M., Stefannelli, M., Magnani, L., Barosi, G.: *An epistemological framework for medical knowledge-base systems*. *IEEE Transactions on Systems, Man, and Cybernetics* 22(6), 1361–1375 (1992)
29. Thagard, P.: *The Best Explanation: Criteria for theory Choice*. *The Journal of Philosophy* 75, 76–92 (1978)
30. Thagard, P.: *Computational philosophy of science*. MIT Press (1988); Trigg, G.I. (ed.) *Encyclopedia of Applied Physics*, vol. 2. VCH Publishers, Inc., New York (1991)
31. Zeilik, M.: *Astronomy: the evolving universe*, 9th edn. Cambridge University Press, New York (2002)

# Moral Intuitions vs. Moral Reasoning. A Philosophical Analysis of the Explanatory Models Intuitionism Relies On

Sara Dellantonio and Remo Job

**Abstract.** The notion of ‘intuition’ is usually contrasted with rational thought, thus motivating a differentiation between two kinds of processes that are supposed to characterize human thinking, i.e. rational and ‘intuitive’ (immediate and non-argumentative) forms of judgment. Recently, the notion of intuition has also played a leading role in cognitive studies on morality with the rise of so-called social intuitionism, according to which people’s moral stances are *culturally driven intuitions* – i.e. they are quick, involuntary and automatic responses driven by culturally and socially acquired principles (see e.g. [42], [41] and [22]). Usually, intuitionism is presented as radically opposed to rationalistic views of morality according to which moral judgments are the outcome of explicit reasoning. In this work we compare two different hypotheses concerning the possible relationship between reasoning and intuition: a ‘*continuist interpretation*’ (maintaining that intuitions and judgments based on reasoning are produced by the same cognitive process) and a ‘*discontinuist interpretation*’ (supporting the view that they are produced by two different cognitive processes). We argue that a continuist interpretation appears more plausible than a discontinuist one and that the concepts of ‘intuition’ and ‘reasoning’ are two facets of the same process which spans from fast, immediate, and certain answer to slow, conscious and elaborate judgments. According to this interpretation, moral judgments are produced by the same kinds of inferences reasoning relies on, i.e. mostly deduction, induction and abduction. Our analysis will show that to opt for a continuist interpretation has many consequences for the way morality is explained from a psychological point of view. Mainly, it challenges the idea of morality

---

Sara Dellantonio · Remo Job

Dipartimento di Scienze della Cognizione e della Formazione,  
Università degli Studi di Trento

e-mail: [sara.dellantonio@unitn.it](mailto:sara.dellantonio@unitn.it), [remo.job@unitn.it](mailto:remo.job@unitn.it)

proposed by intuitionism, according to which moral intuitions are rigidly driven by culturally learned principles.

Our reflections lead rather to the conclusion that the first and spontaneous intuitions fully enculturated people may experience do not often express the best moral judgment possible in a certain situation, but are rather the product of the prejudices people inherit from their culture/subculture. This gives rise to the conclusion that people are better guaranteed to form truly moral judgments when they do not respond intuitively to morally relevant situations, but interrupt and override this automatic processing, moving on to a controlled i.e. a rational process.

## 1 Introduction

The notion of ‘intuition’ has continued to be influential in the philosophical tradition since the pre-Socratics. Over time, however, it has evolved taking on deeply different connotations. In contemporary philosophical studies, intuition is viewed as an immediate, simple, passive, non-verbal procedure of knowledge acquisition (see e. g. [43]). In cognitive science, this notion is usually contrasted with rational thought, thus motivating a differentiation between two kinds of processes that are supposed to characterize human thinking, i.e. rational and ‘intuitive’ (i.e. immediate and non-argumentative) forms of judgment<sup>1</sup>. Recently, intuition has also played a leading role in cognitive studies on morality and moral sense, since it is considered an ideal concept to describe the way in which people produce their moral judgments.

Within this dualistic view of moral judgment, it has been proposed that people’s moral stances are *culturally driven intuitions* – i.e. that they consist of quick, involuntary and automatic responses driven by culturally and socially acquired principles (see e.g. [42], [41] and [22]) – and that these intuitions are the product of an innately programmed moral module in the brain (see e.g. [28] and [30]). More precisely, intuitions are defined as “the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring conclusions. [...] One sees or hears about a social event and one instantly feels approval or disapproval.” ([22] p. 818). At least according to Haidt’s social form of intuitionism, the good/bad evaluations produced by moral intuitions “are made with respect to a set of virtues held to be obligatory by the culture or the subculture.” ([22] p. 817)

Usually, intuitionism is presented as radically opposed to the previous rationalistic view of morality (see e.g. [22], [23] and [24]) supported by a large part of the classical philosophical studies and by the psychological tradition that starts with Piaget and continues with Kohlberg and Turiel (see e.g. [47], [35] and [55]), according to which moral judgments are the outcome of explicit

<sup>1</sup> For an overview see e.g. [9].

reasoning. Such reasoning is considered to be a form of conscious reflection or of verbalized deliberation that proceeds slowly and with effort, weighing up motives and principles.

Intuitionism tries to account for the fact that in order to form a moral judgment people often do not reason about an issue or weigh up the different aspects of a situation. Rather, their answer regarding the right thing to do seems to come up immediately and spontaneously, and its content tends to conform to the rules and the habits of the culture or group they belong to. Indeed, this same evidence constitutes the starting point of many contemporary cognitive theories about morality (like e.g. the Rawlsian and Humean ones: see [4]). In this sense, intuitionism is surely right in saying that, to be plausible, moral theories need to explain why moral judgments appear (at least mostly) to be intuitive rather than reflective. Still, we think that the concept of ‘intuition’ which intuitionism appeals to hasn’t been defined precisely enough from the point of view of the cognitive processes that are supposed to produce intuitions. The following questions need at least to be investigated: What kind of process gives rise to intuitions? In what respect does this cognitive process differ from the one that it is supposed to produce reasoning?

In this work we compare two different hypotheses concerning the possible relationship between reasoning and intuition. I.) On the one hand, we consider the hypothesis that our intuitions (i.e. the fast and immediate answers people produce in certain cases, without having doubts or being aware of the reasons supporting them) and reasoning (the slow, reflective and often beset by doubts form of thought people sometimes perform) are produced by the same cognitive process, using the same kind of information. We will call this a ‘*continuist interpretation*’ of the relationship between rationalism and intuitionism. II) On the other hand, we will consider the idea that intuitions and judgments based on reasoning are produced by two different kinds of cognitive processes and therefore really do differ cognitively from each other, as assumed by intuitionists. We will call this a ‘*discontinuist interpretation*’ of the relationship between rationalism and intuitionism. On the basis of this comparison we will argue that a continuist interpretation appears more plausible than a discontinuist one and that the concepts of ‘intuition’ and of ‘reasoning’ do not cognitively differ, i.e. they do not refer to the outputs of two different cognitive processes, but are two facets of the same process which spans from fast, immediate and certain answer and to slow, conscious and elaborate judgments. According to this interpretation, moral judgments are produced by the same kinds of inferences reasoning relies on, i.e. mostly deduction, induction and abduction.

This thesis concerns only moral judgments. In this sense it is not a claim against dual theories or the massive modularity view outright, according to which the mind works using two radically different systems or processing mechanisms: a modular system, which works rapidly and automatically and

a non modular system that produces complex, hypothetical and decontextualized thought, which is flexible and able to reach high abstraction levels (see e.g. [10], [48], [7] and [53]). However, *as far as moral cognition is concerned*, the hypothesis we put forward here is surely incompatible with dual-*system*-theories, but still compatible with more recent dual-*processing*-theories that differentiate between various “types” or “levels” of processing, leaving open the possibility that they might be generated by the same system ([8] and [16]).

Our analysis will show that to opt for a continuist interpretation has many consequences in terms of the way morality is explained from a psychological point of view. Mainly, it challenges the idea of morality proposed by intuitionism, according to which moral intuitions are rigidly driven by culturally learned principles, and to be morally virtuous simply means to be “fully enculturated”, i.e. to have assimilated the moral principles of a culture or subculture and to follow them slavishly. Our reflections lead rather to the idea that the first and spontaneous intuitions fully enculturated people may experience do not often express the best moral judgment possible in a certain situation, but are rather the product of the prejudices people inherit from their culture (or subculture). A parallelism with socio-psychological studies on this aspect is proposed. This gives rise to the conclusion that people are better guaranteed to express truly moral judgments when they do not respond intuitively to morally relevant situations, but rather interrupt and override this automatic processing, moving on to a controlled i.e. a rational process.

## 2 Intuitions and Inferential Reasoning According to a Continuist Interpretation

Apparently, the distinction between moral reasoning and moral intuition is clear and sharp, even easy to observe in our everyday experience. And apparently intuitionism is right in maintaining that our ‘moral’ judgments seem to be, at least in the large majority of cases, entirely ‘intuitive’, since they are fast and since they are not beset by doubts, while in just a few particularly difficult situations, involving different and possibly contradictory aspects (like the moral dilemmas made up in the laboratory using artificial scenarios) people have recourse to a slow and reflective form of reasoning in order to form moral judgments. Still, the fact that our moral judgments are experienced as being mostly intuitive and that intuitions are experienced as cognitive processes which radically differ from reasoning does not guarantee that there is an actual difference *in the nature of the cognitive process* that produces what we perceive to be an intuition and what we perceive to be a form of reasoning. In fact, in this respect questions arise about what kind of cognitive process may produce moral intuitions and whether this process is really qualitatively different from the one underlying reasoning.

To follow up this line of investigation we need firstly to consider whether it is possible to explain intuitions and judgments based on reasoning as two apparently different products of the same cognitive process. We will call this a ‘*continuist interpretation*’ of the relationship between intuitions and judgments based on reasoning.

Intuitions and judgments based on reasoning are usually considered to differ from each other first of all because intuitions are produced without people being conscious of the possible reasons supporting the judgments they intuitively formed. However, the unconscious nature of the cognitive processes that lead to an output is not a distinctive feature of moral intuitions. In fact, the idea that the way we form our thoughts (the information we use and the steps we follow) is in general for the most part not accessible to consciousness is one of the essential tenets of the cognitive sciences, which states that cognition consists in information processing, of which only the final product is accessible to consciousness. According to a classic computational view, thought is produced by a central system that processes information on the basis of logical and inferential relations that refer to the semantic properties of the information processed. As Fodor makes clear “[...] the notion of computation is intrinsically connected to such semantic concepts as implication, confirmation, and logical consequence. Specifically, a computation is a transformation of representations which respects these sorts of semantic relations.” ([13] p. 5) According to this view, when I hear for instance a sentence like ‘Cleo is lying on the floor’, my immediate understanding of it and my automatic reaction to her – I run towards Cleo – will depend on information processing which will follow more or less a path such as:

Cleo is a fish  $\Rightarrow$  Outside the water fishes die  $\Rightarrow$  On the floor there isn’t any water  $\Rightarrow$  Either Cleo has already died or she will soon, unless I immediately put her back in the water.

Even though this information processing is the condition for understanding the sentence I hear and its consequences, I don’t need to be conscious of the path it followed to understand the sentence and to react to it. In fact, people are generally not conscious that they are processing information in this way and they may not even be able to reconstruct the information process through which they came to understand a sentence when they are requested to explain it. Even if they are able to do so, it will cost them a lot of effort to make explicit and verbalize linguistically the inferential path that produced the understanding. And in any case the explanation of the inferential path is just a *post hoc* reconstruction; and one can never be sure that the reconstruction corresponds to the actual inferential path that has taken place. *As for the understanding of the sentence itself it will appear to the subject as an unconscious, immediate, spontaneous and non-reflective intuition.*

What this example shows is that we can reach a specific conclusion (the understanding of something, but also a certain judgment) on the basis of inferential processing on semantically structured information *without being*



*conscious that this information is being processed.* This processing may also start spontaneously, be fast and not require any particular reflection. But, if so, then both our intuitive (fast, spontaneous, unconscious) answers and our reasoned (slow, conscious) judgments could be the results of information processing which relies on logical and inferential operations on semantically structured information. This semantically structured information could be that which concepts are composed of (see also [4]).

Indeed, according to mainstream research on semantics (contra Fodor's atomism: see e.g. [14], [15]) concepts are composed of different pieces of information: According to this view, to know what a fish is – i.e. to have the concept of 'fish' – means for example to know (at least) that fishes are animals, that they can live only in water, that they have round open eyes, fins, and commonly a typical rounded-stretched form, that they don't have legs, etc; the concept of 'fish' must therefore be made up of these 'pieces of information' (in the literature on concepts they are more often called 'features'<sup>2</sup>). These features do not codify perceptual information only, i.e. information about the external aspect of the conceptualized things; they also codify the common, widely shared and well-grounded knowledge people have about objects: someone who knows e.g. what a fish is (i.e. who has the concept of fish), generally knows a lot of things about fishes such as: they are mostly edible, some of them are considered pets, they eat insects, lay eggs, etc. Thus, this information is also part of the concept of fish.

To claim that all kinds of judgments are the results of information processing amounts to stating that all kinds of judgments – intuitive or reasoned – are the results of information processing which relies on logical and inferential operations on concepts and on these concepts' features. A general idea about how this might work can be given using the example above: since the concept of 'fish' includes the information 'animal' and 'animal' includes the information 'mortal', one can infer that – being an animal – a fish is mortal. In general, the semantically structured pieces of information that compose concepts can be connected with each other to form chains of deductive and inductive inferences as well as other more complicated forms of inferences that we use in our reasoning processes like abductive inferences, which are indispensable for forming hypotheses.

Smith states that: "An intimate relation connects inductive inferences and categorization; namely, categorizing an object licenses inductive inferences about that object. For example, if we see a round, reddish object on a tree and categorize it as an apple, we can then infer that is edible and has seeds. Thus categorization is the mental means we have for inferring invisible properties from visible ones." ([50] p.6) Even though Smith focuses primarily on inductive inferences, the categorization process as he describes it also involves deduction and abduction. If I see a round, reddish object on a tree and all the round, reddish objects I have seen on trees in the past turned out to be

---

<sup>2</sup> For a technical overview of the featural approach see e.g. [50], pp. 10-22.

apples, I can inductively infer that this round, reddish object is an apple too. But, if I know that all apples are edible and have seeds, then I can deductively infer that, if this is an apple, it will be edible and have seeds. As far as abduction is concerned,<sup>3</sup> this is used to form hypothesis about the objects we have categorized (about their behavior and their connections with others). Referring to the example, abduction is used in cases like:

Cleo is on the floor  $\Rightarrow$  therefore, she has jumped out of her bowl.

What happens from a cognitive point of view, when we make this inference is something like:

*observation*: Cleo (my goldfish) is on the floor  $\Rightarrow$  *surprising effect*: Cleo should not be on the floor, she should be in her bowl, where I left her  $\Rightarrow$  elaboration of an hypothesis to explain the event: (fishes can jump) if Cleo had jumped out of her bowl, this would explain why she is now on the floor; (given what I know about fishes and about other circumstances regarding the environment where the bowl was located) no other hypothesis can explain the event as well as this one  $\Rightarrow$  *conclusion*: hence, Cleo must have jumped out of her bowl.

What this example suggests is first of all that there is an intimate relation between concepts and inferential reasoning in all its forms; more precisely: that all forms of thought based on concepts work on logical and inferential operations. In this sense, all forms of thought based on concepts are in a way rational, if by rational we simply mean ‘based on logical and inferential operations’ applied to available information’.

This conclusion allows us to now specify more precisely the position expressed by a ‘*continuist interpretation*’ of the relationship between reasoning and intuitions: according to a continuist interpretation, both (fast and automatic) intuitions and (conscious) reasoning are forms of thought based on concepts – i.e. they are products of logical and inferential operations on concepts and on the pieces of information concepts consist of; for this reason and in this respect they are both in a sense rational.

According to this interpretation, the word ‘rationality’ does not describe conscious forms of information processing only; nor does this notion of rationality include any guarantee that the conclusion of an inference will be ‘rational’ in the sense of ‘the best possible’ all things considered. In fact, the logical and inferential operations on concepts we are speaking about are based just on the specific pieces of information a person has acquired about the world and is able to include in a specific occurrence of information processing and this information is often very limited, inaccurate, reciprocally incoherent, and most importantly *oversimplified* and *affected by prejudices*. (We will say more about this later.)

Furthermore, the idea that all forms of thought are inferential information processing based on logical and inferential operations (i.e. that they are in

<sup>3</sup> For a classical definition of abduction see e.g. Peirce [44] §188-189; for a contemporary discussion on abductive inferences see e.g. [58] ch. 1; [38] ch. 1, 2.

a way rational) is not in conflict with the possibility that emotions play a part in processing information. Even the simplified example of the sentence 'Cleo is lying on the floor' shows on the contrary that emotions are always part of information processing, since the chain of thoughts that follow the understanding of this sentence as well as our practical reaction to it are both highly dependent on whether and how much we care that Cleo may be dying. So, it is obvious that information isn't 'emotionally neutral' for a human information processing system, which manifestly exhibits a lot of complex 'positive and negative propensities' towards specific things (it cares/doesn't care for specific things, it likes/doesn't like, it fears/desires etc. certain others). In fact, this particular (positive or negative) 'emotional attachment' to specific pieces of information characterizes all forms of thought (judgments, opinions, decisions). It is the particular form of this attachment that drives e.g. our reaction to the sentence 'Cleo is lying on the floor': a positive attachment to Cleo makes us run to save her, while a negative one makes us wait a little longer. This applies to all forms of thought/opinion/conclusion from the simplest ones such as 'The bus is leaving in five minutes' (Do we care? How much do we care? Is it worth running and e.g. giving up our morning coffee?) to the most socially complex ones like 'In some places children starve to death' (Do we care? How much do we care? Is it worth giving up some of our income to help them?). In this sense we can consider *all kinds* of inferential processes as driven not only by logical and inferential relations among pieces of information but also by the specific emotional connotations of specific pieces of information.

This description of the inferential processes at the basis of our thinking raises a question: why couldn't moral intuitions just be a form of thought, i.e. why couldn't they be the conclusion of a inferential process like the one that leads to the understanding of and appropriate reaction to the sentence 'Cleo is lying on the floor'? And, if we rely on a weak notion of rationality, why couldn't intuitions and judgments based on reasoning both be realized by the same procedure of the kind just described? In order to support such a continuist interpretation of intuition and reasoning, we need to explain why people experience a difference between these two kinds of process, the one being fast, immediate, sure and unconscious, and the other being slow, reflective and beset by doubts, even though both were produced by the same mechanism.

One possible answer to this is that intuitions and judgments based on reasoning are experienced as different cognitive modalities because they are the expression of two possible courses of the supposed inferential information processing. When the information process proceeds without 'hitches' – i.e. when the logical and inferential operations on the available information go on without running into a contradiction or obstacle of some kind and do not encounter any novel or surprising situation that needs to be weighed up carefully – their results appear to our consciousness in the form of quick, immediate, and spontaneous intuitions. In contrast, when the information

process encounters an impasse, (which may also be caused by confrontations with other people and/or by the need to find explicit arguments in support of a position), then it takes the form of reasoning – i.e. a reflective, conscious, slow and difficult form of thinking. A soldier may, for example, come intuitively to the conclusion that it is morally permissible to torture his/her prisoner to draw information out of him/her. Still, if he also happens to conceive of his/her prisoner as a person fighting for what he/she believes in and for his/her people, the soldier's moral position may reach an impasse which needs to be solved. In sum, finding a way to restore the coherence of the system is a necessary condition in order to arrive at a judgment about the moral legitimacy of torturing the prisoner.

So, to sum up, according to a continuist interpretation, intuitions and judgments based on reasoning could be produced by the same inferential process, while the difference the subject experiences between them could be due to the fact that inferential processes may take different paths: when the process does not encounter any obstacles, subjects experience the conclusion as an intuition, when, on the other hand, the process does meet an obstacle, coherence needs to be restored, the process slows down, different possibilities are explored, and sometimes new information is collected. In this case the subject experiences the conclusion of the process as reasoned.

### 3 The Phenomenon of 'Dissidence'

The explanation given in the previous section opens the door to the possibility of a continuist interpretation of intuitions and judgments based on reasoning, according to which they are produced by the same cognitive mechanism which sometimes proceeds without 'hitches' and appears to be fast, immediate, sure and unconscious, while at other times encounters an impasse or runs into a contradiction of some kind or encounters a novel or surprising situation and takes the form of a conscious, slow and difficult form of thinking. However, the possibility of arguing for a continuist interpretation of intuitions and judgments based on reasoning does not *ipso facto* exclude that a 'discontinuist interpretation' is also plausible and that authors like Haidt who favor an account based on intuitions are right in maintaining that intuitions are radically different from reasoning. According to such an interpretation, intuitions *are not* produced by the same cognitive process as reasoning, rather they are produced by an automatic and less flexible (modular) mechanism than reasoning which works just with a specific type of information, i.e. with moral information.

Haidt maintains that this mechanism is set by the moral principles of the cultural group or subgroup people are part of, while moral intuitions are directly triggered by the mechanism itself (see e.g. [22] and [28]). So, drawing some examples from Haidt, if a culture sets the individuals' mechanism according to the principles 'incest is always forbidden' or 'abortion is always

forbidden', people belonging to that culture will always have the intuition that abortion and incest are morally forbidden in any case. This intuition will be immediate, unequivocal and unquestionable and will come up in the form of a strong feeling of right or wrong. In fact, in consequence of this interpretation, Haidt states that moral virtue and full cultural integration are one and the same thing: "a fully enculturated person is a virtuous person" ([24] p. 216). So, once the modular mechanism is set up by the culture, its working will be strongly bound to the principles it works with and leave very little space for change and flexibility.

A discontinuist interpretation clearly more closely adheres to the intuitionists' point of view since it accounts for the idea that moral reasoning and moral intuitions are not only experienced as different kinds of judgments, but they actually are different kinds of judgment since they are produced by two different processing systems. While reasoning is produced by a flexible system that can make use of all kinds of information, carry out all kinds of inferences, reflect and draw conclusions, intuitions are rigidly driven by culturally learned principles. The point of intuitionism is that, when we form or express a moral position, we do so not on the basis of reasoning, but rather on the basis of intuition. Still, this interpretation runs into difficult problems. A first and fundamental one is that it gives rise to a concept of morality that does not correspond to what have always been considered truly moral attitudes and stances.

In fact, morality cannot consist of a supine allegiance to the norms and customs of a group. As both the classic philosophical and psychological tradition have shown, morality cannot merely consist in blindly following a rule, without evaluating whether this rule is morally right or not (see e.g. [33] for the philosophical tradition and [35] for the psychological one). On the contrary, truly moral forms of thought and behavior are those which are capable of breaking away from the norms and the customs of a particular group in order to follow different principles, which are considered as right independently of what it is stated by the group or sub-group one belongs to. This behavior has indeed been placed by Kohlberg at the 5<sup>o</sup> and higher level of his moral developmental scale (see e.g. [35]) and is supposed to be based on the individual's capacity to critically and autonomously evaluate the right moral behavior in a given situation. From this perspective, one of the moral conditions *par excellence* is the phenomenon of *dissidence*, i.e. a form of disagreement expressed at a certain point by a member of a group about a principle, or about a position belonging to the common ideological framework of the group (see also [4]). Among the most well known examples of dissidence is the case of Nazism and of those German Aryanists that adhered, at least at the beginning, to National Socialism, but later helped Jews to save themselves, betraying in so doing the ideals of their group and infringing the racial law in force.

The fact that this phenomenon poses a problem for social intuitionism has already been pointed out very clearly by Nervaes, who uses Kohlberg's

position to make a critical point against Haidt and Bjorklund: “In the early years of the moral developmental tradition, there was a distinction made between social conformity and moral development (35). The distinction was necessary in order to explain how in some situations (e.g. Germany in the 1930s) social conformity worked against moral development, and in others resisting social pressure (U.S. civil rights movement of the 1950s and 1960s) was a virtuous path. Thus, it is shocking to read Haidt and Bjorklund assert that ‘a fully enculturated person is a virtuous person’ (24 p. 216). Apparently Hitler youth and Pol Pot’s Khmer Rouge were virtuous and most moral exemplars are not.” (40 p. 239) And further: “[...] how does social intuitionist theory judge the goodness or badness of particular intuitions? Intuitions appear to be equally meritorious, as are all cultural practices, if they conform with the norms of one’s social group (‘full enculturation’). This is precisely the attitude that drove Kohlberg to mount his research program – how to support the law-breaking behavior of Martin Luther King, Jr., and condemn the law-abiding behavior of the Nazi soldier.” (40 p. 240)

Nervaez’s objection applies to all views that, like social intuitionism, consider morality as the output of an automatic cognitive process, which is uniform for all people of the same group and which is driven by the moral principles sanctioned by that group. The problem with such views is that they do not account for the autonomy of moral positions with respect to the moral principles accepted and shared by a group or sub-group. A theory of moral cognition may take the position that dissidence isn’t an emblematic expression of moral behavior. Still, since the phenomenon of dissidence constantly occurs in history and since it has always been considered as a genuine moral stance both by the people who took a dissident position and by the people witnessing the situation from a point of view external to the group, it needs to be accounted for by a theory of moral cognition aiming to provide a comprehensive explanation of the human ‘moral sense’. The phenomenon of dissidence suggests that humans have the capacity to morally act in a way that infringes the moral principles and conventions embraced by the community, group or sub-group they belong to. This means that people do not merely follow the moral principles embraced by their group, but they are also able to identify, work out and weigh up *critically* and *autonomously* moral principles and moral behaviors. For this reason they may arrive at a judgment that diverges from, or is opposed to, the one expressed by the norms of the customs of their group.

According to theories that assume moral judgments are intuitions driven by social principles, all kinds of traceable differences among moral intuitions can only be ascribed to cultural differences, or more precisely to more or less fine-drawn differences among the principles people happened to learn during their life. There isn’t any reason in principle to exclude that the phenomenon of dissidence can be explained in this same vein as the consequence of the fact that different people are ‘exposed’ to different cultural information or have ‘assimilated’ different cultural elements. Still, in order for this proposal to

hold, intuitionist theories need to clarify *why* and *how* this may happen: i.e. what does this different ‘exposure’ and ‘assimilation’ concretely consist of and which kinds of information among the varieties available are relevant to direct subjective moral intuitions in one direction or in another. The problem here, of course, is that, since each of us belongs concurrently to different groups and subgroups, allowing very subtle idiosyncratic cultural differences to affect our moral judgment, we must admit that each of us is determined by his/her own unique cultural experience. But, in this case, the notion of “enculturation” would become explanatorily useless.

In addition, such an explanation will be difficult to sustain on a theoretical level if we adopt a discontinuist interpretation of intuitionism. According to such an interpretation intuitions aren’t the product of the central system, but of a module, i.e. by definition a mechanism which is much more inflexible and informationally encapsulated than the central system and which can hardly rearrange itself and become sensitive to new information. Such a mechanism must therefore be almost insensible to new information acquired by the system after the time when it is first set (encapsulation). This mechanism must also be quite resistant against distortions (i.e. untouched in its *modus operandi*) brought about by new information. These characteristics make it particularly difficult to explain cases like the phenomenon of dissidence in which moral judgment changes radically over time. For such a change to happen the modular mechanism for the production of moral judgments must be both quite permeable to new information (even to information which is opposed to specific culturally dominant moral principles) and quite flexible in order to turn the old operational mode into a new one and find a new assessment after assimilating new information. If we take perception as an emblematic example of cognition produced by modular mechanisms – as is usually done, and as Haidt does as well (see e.g. [22] p. 814) – we can easily face the problem with flexibility and encapsulation of modules: the way we perceive neither changes over time when we acquire new information, nor is it influenced by information other than that specific information needed to first set up the mechanism and which the mechanism has access to.

Haidt admits that inflexibility and encapsulation pose a problem for a theory of moral cognition and maintains that intuitionism needs for this reason to rely on a weaker idea of modularity like the one proposed by the so called ‘massive modularity hypothesis’ (p.es. [27], [28]). Still, even if we give up completely or to a large extent the idea that modules have the properties of being rigid and encapsulated (at the risk, by the way, of making the modularity thesis lose its sense, since the supposed modules could become identical to the central system), intuitionism – i.e. the thesis of morality as full enculturation – does not allow us to explain why someone can become a dissident even though he/she is and continues to be part of a group/sub-group that upholds different moral values. In other words, interpreted according to a discontinuist interpretation, intuitionism cannot explain why the moral

judgment of a person can change without any correspondent modification in the cultural environment he/she is exposed to.

## 4 The Slow Processing of Morality

Why and how might a person change his/her moral judgment over time, infringing the cultural principles he/she first learned? In the previous section we tried to show that a continuist interpretation of intuitionism and reasoning provides us with better theoretical means to answer this question, because it suggests that, if necessary – i.e. when the situation is perceived as novel and or presents obstacles, impasses or contradictions – people may modify the way they produce their moral judgments. They can shift from intuitions to slow and conscious reasoning, adding and weighing more and more elements in their inferential processing. According to a continuist approach, the inferential apparatus deployed in producing moral judgments is the same whether a fast or a slow response is provided.

Both the continuist and discontinuist interpretations are compatible with the idea that the fast and intuitive way of processing a moral output is cognitively realized using what in the literature on rationality and decision making are called “heuristics”, i.e. specific procedures that speed up thinking processes allowing a parsimonious search for information and giving rise to immediate and spontaneous answers that are perceived as intuitions.<sup>4</sup> However, the idea of what a heuristic is and above all what effects the application of a heuristic has on the processes that lead to the production of a moral judgment radically differ in the two cases.

Because heuristics ignore potentially relevant information, they have always been considered error-prone and less-than-optimal procedures (see e.g. [31] and [32]). Still, recent studies have shown that heuristics may be adaptive and ecologically useful (see e.g. [19] and [20]) hence suggesting that they may be, at least in some cases, preferable to reasoning, including moral reasoning. Indeed, some authors have implicitly or explicitly maintained that there is nothing wrong with the use of heuristics to produce moral judgments. This has been implicitly assumed by Haidt when he states that moral evaluations are (and can only be) intuitions driven by cultural values (see e.g. [22] and [24]), and it has been explicitly put forward by Gigerenzer, in discussing Haidt’s position ([18], pp. 18ff). While it might be the case that for decision making intuitions-as-heuristics often provide positives outcomes, for moral

---

<sup>4</sup> Some authors explicitly connect the idea that moral judgments might be produced on the basis of heuristics with social intuitionism; it is e.g. Gigerenzer who says: “[...] moral intuitions as described in the social intuitionist theory (e.g. [22]) can be explicated in terms of fast and frugal heuristics” ([18] p. 9; see also [17]). However, this idea is equally or possibly even more compatible with a continuist approach, according to which heuristics are just a procedure applied by the central system to speed up its processes.



dilemmas they may be misleading since they might not be *the result of some cultural moral principle but, rather, of some cultural prejudice*.

A case that may help in clarifying the risk of applying intuitions when we are requested to evaluate a person or a situation from a moral point of view is that of ‘stereotypes’ as they are defined by social psychology as a form of heuristic used by people to speed up their judgments about a social situation. Social psychologists define ‘stereotypes’ as “knowledge structures” that people use to categorize groups or specific members of groups ([52] pp. 1-8). Such knowledge structures tend to work on the basis of a limited number of attributes and to disregard individual differences, leading to unwarranted generalizations about individuals and groups. Just like heuristics, stereotypes simplify information processing (see e.g. [1], [12], [36], [37] and [34]) by reducing variability in the input and by tracing something newly experienced back to already available knowledge structures. Furthermore, as in the case of heuristics, the activation of stereotypes occurs quickly, automatically, spontaneously and effortlessly, without intention or consciousness, when we first categorize a person as a member of the group we have stereotyped (see e.g. [2] and [57]).

Stereotypes have a strong cultural base, and are powerful tools for processing in-group/out-group relationships. For this reason, they tend to associate positive properties with the members of one’s own group(s) and negative properties with the members of other groups. It is such cultural filtering of information that make stereotypes special cases of intuitions: They have played and may still play an adaptive role by preserving in-group safety, but they cannot be taken as morally positive stances since individuals are judged not on the basis of what they are, but on the basis of what the prejudices about the group they belong to suggests they may be.

For instance, we could intuitively judge that it is morally legitimate to restrain gypsies from moving freely from country to country because they might rob other people’s properties. But in so doing we are ascribing to each individual gypsy the feature ‘thief’ that may apply to some of the group members. Conversely, we may suspect a gypsy of robbing something as the result of the prejudice that – since he is a gypsy – he must surely be the person responsible for robbery (if not presently, at least in some other cases) and therefore he deserves to be punished. Thus, independently from any consideration regarding whether the use of heuristics as short cuts in replacement for longer reasoning processes have to be considered adaptive or ecologically useful as regards the reaction or interaction they elicit toward the categorized instances, from a moral point of view they cannot be viewed as optimal. And their sub-optimality, or plain wrongness, is due to the fact that they are culturally filtered, i.e. they are the product of people’s ‘full enculturation’, and therefore of the cultural prejudices people may have. From this point of view, while intuitions may still give rise to correct moral judgment, the very fact that they rely on limited conceptual information and, as in the case of

stereotypes, culturally filtered information, makes them less reliable as far as the output of the moral judgment is concerned.

According to the discontinuist interpretation, the moral positions people express are produced by an intuitive, modular system specialized for the processing of moral outputs and separated from the central system that carries out reasoning processes. According to this view, the reasoning process starts only after the moral module has produced its output and therefore cannot influence or intervene in the work of the moral module or change its output, but only deliver a post-hoc justification of the output produced by the module, whatever this output may be. In this approach the question about what is morally right or wrong can only be decided on the basis of intuitions and never on the basis of reasoning. So, from this approach it follows that the intuitions we form on the basis of our prejudices are the only moral positions we are able to produce: people don't have an alternative to making moral judgments on the basis of their own prejudices. In the continuist as contrasted with the discontinuist interpretation, this view on moral judgment changes completely because it is no longer assumed that the production of moral judgment is the exclusive prerogative of the intuitive system. In fact, by hypothesizing that intuitions and reasoning are not produced by two separate, independent systems, but rather by a single process that can proceed either quickly, using a limited amount of information, or slowly including more information, we conclude that both what we call reasoning and what we call intuitions can produce moral judgments.

Besides being highly undesirable, the conclusion of the discontinuist approach is also not plausible because we are *de facto* at least potentially capable of escaping our own prejudices and producing moral evaluations based on complex information, even when stereotypes are available. In fact, the example of stereotypes allows us to point out that a person can voluntarily generate obstacles when trying to escape his/her own prejudices. As some results obtained in the field of social psychology show, if people are motivated to challenge their own prejudices, reasoning might contribute to identifying them, to discovering that they are at work, and to reducing their effect on a final judgment.<sup>5</sup> Such a possibility was pointed out by Allport as early as

---

<sup>5</sup> Social psychological models make various hypotheses about the relation between intuitions produced immediately and spontaneously on the basis of our prejudices and reasoning, which can be activated consciously to overcome these prejudices. Some models tend to stress the duality of judgment processes, suggesting that, when someone weighs up a situation or a person, he/she can give either an intuitive evaluation or a reasoned judgment, (see e.g. [45]), while other models assume that the two type of processes occur in parallel and can affect each other (see e.g. [51]). We suggest that in reality both situations can occur and can be descriptively appropriate, since we sometime evaluate a situation only intuitively or only on the basis of a reasoning process, while at other times we spontaneously give an intuitive response to a situation, while concurrently activating a reasoning process about it.

1954 in *The Nature of Prejudice*, where he maintains that people can “put the brackets on their prejudices” ([1] p. 332). More recently the possibility of escaping prejudice has been considered more diffusely (e.g. [11] and [5]). In particular, it has been shown that people are able to intentionally inhibit stereotypes, and the influence of these stereotypes on judgments, and to replace them with other kinds of knowledge on the matter ([5]). Further, that this control cannot be engaged without becoming aware of the presence of a prejudice: i.e. without a conscious reflection concerning the fact that a bias is at play (see e.g. [3], [6]). The possibility pointed out by social psychology of escaping from our first prejudicial intuition turns out to be compatible with a continuist interpretation. According to this view, we usually produce our moral evaluations using fast and frugal heuristics, but when we have grounds to avoid short cuts and easy solutions, we are able to slow down the process and to reason using a larger amount of information.

## 5 Morality, Culture and Educational Level

Haidt’s orthodox discontinuist interpretation suggests that the truly moral responses are intuitions, whose content is entirely determined by the cultural principles and values peoples have assimilated. As a consequence of this view people of the same cultural group share a large and strong intersubjective agreement with respect to their moral positions, while the contents of the moral positions of different groups may differ greatly from each other. However, as we will try to show here, this idea of *a wide in-group moral uniformity, accompanied by a wide trans-group moral dissimilarity* conflicts with some phenomena pointed out by moral psychology with relation to human moral responses. In particular, we will present some data driven by moral literature that, taken together, indirectly support our view, at least since they show that an increase in people’s information– i.e. an increase in the features that qualify the concepts they rely on in their reasoning processes<sup>6</sup> – changes people’s moral stances, decreasing the influence of their culture and forming something like a trans-group similarity among moral judgments.

---

<sup>6</sup> The notion of ‘information’ is used here in a purely cognitive sense to mean the features our concepts and conceptions are made of. This notion has nothing in common with that of ‘information systems’ or of ‘information made available by the media’, since – in opposition with some philosophical views like Habermas’ – we don’t think that an increase of the information made available in a society by the media would lead to a more moral and democratic system (see e.g. [21]). Cognitive capacities of humans are indeed very limited and are usually focused on specific tasks or aspects. An increase in the general information available could just be ignored, or not be correctly assimilated or lead to confusion and misunderstandings. On the contrary, when we appeal to a form of reasoning that applies to the information our concepts and conceptions are made of, we are speaking of information already available to the subjects, whose features can become explicit for him/her.

The best way to introduce this point is in terms of the contraposition of Turiel's and Heid's views on morality. Throughout the course of his research Turiel has been trying to show that human beings are equipped with the basic capacity to distinguish moral violations from merely conventional violations. According to Turiel's definition: "Conventions are part of constitutive systems and are shared behaviors (uniformities, rules) whose meanings are defined by the constituted system in which they are embedded" while moral rules are "unconditionally obligatory, generalizable and impersonal insofar as they stem from concepts of welfare, justice, and rights" ([56] p. 169-170).

Turiel and the other authors that have investigated this position (see e.g. [49] and [54] for reviews) point out that moral violations are perceived (both by children and adults) as more serious than the conventional ones; as independent from any authority that imposes them (like parents, teachers, governments or even God); and as ubiquitous, i.e. as not bound to any particular place, context, culture or habit. So, according to this characterization, while an act like e.g. eating with the hands is perceived as a conventional violation, which is not very serious and applies only in some places, contexts and cultures but not in others, and which depends on an authority, an act of violence is perceived as a serious moral violation, that applies everywhere and does not depend on any authority that imposes a restriction or compliance.

Nevertheless, the idea that it is possible to trace an univocal, transcultural and unanimous difference between conventional norms and moral norms on the basis of the criterion that the only properly moral violations are those related to welfare, justice, and rights has been challenged by a different research tradition, lead by Haidt. Haidt's research shows in fact that what people recognize as properly moral violations depends on both social and cultural factors.

As far as social factors are concerned, Haidt's experiments show that the moral intuitions of people are deeply influenced by their socioeconomic status, therefore, indirectly, by their level of education, and more generally, by the variety of their contacts and by the amount and quality of their experiences and knowledge (see e.g. [29]). According to Haidt, theories relying on Turiel's position, which narrow the moral domain to issues of harm/care and fairness/reciprocity/justice are 'parochial' and biased by the fact that the researchers carrying out these studies are more often liberal, and investigate therefore only the moral values they recognize as such. In addition, the data collected by these researchers are further biased by the fact that experimental subjects typically come from the same social group, since they are mostly university students and colleagues. Haidt shows that while well-educated people with a high socioeconomic status, especially secular and liberal Westerners, do actually identify the domain of morality only with phenomena related to welfare, justice, and rights, people of low socio-economic status consider as properly moral also other kinds of violations involving things which are offensive, disrespectful or disgusting (as e.g. having sex with a chicken carcass or cleaning the toilet with the National flag: see [29]).

It is to explain these aspects that Haidt appeals to the cultural factors that influence moral intuitions. Further studies by Haidt carried out on cultures other than the ones usually considered in academic research – i.e. studies investigating cultures other than the North American and European ones, or also addressed to conservatives in Western cultures – show that people may also consider as properly moral (and not just as conventional) issues of in-group/loyalty, authority/respect and purity/sanctity (see e.g. [25], [28] and [26]). As Haidt points out also relating his point to the empirical research of other authors: “in most cultures the social order is a moral order, and rules about clothing, gender roles, food, and forms of address are profoundly moral issues. [...] In many cultures the social order is a sacred order as well.” ([28] p. 371) And further: “[...] only an elite American college population limited the moral domain to matters of harm, rights, and justice. For other groups, particularly for low socioeconomic status groups in Brazil and in the United States, actions that were disrespectful or disgusting were said to be morally wrong (universally wrong and unchangeable) even when respondents specifically stated that nobody was harmed by the action.” ([28] p. 372)

If we cross Haidt’s investigations with the conclusions reached by Turiel we achieve a result which is as unsurprising as it is interesting and difficult to explain on the basis of Haidt’s theory. (a) Firstly, although not all cultures or groups restrict the domain of morality to issues related solely to harm/care, fairness/reciprocity/justice, these nevertheless represent something like a ‘lowest common denominator’ or ‘hard core’ of moral cognition, *which is shared by everyone* (i.e. by people of any origin, educated or not, liberal or conservative, religious or secular, belonging to one culture or to another). The idea of a lowest common denominator of moral sense focused on harm-fairness-based violations is also shared by other studies, as e.g. the ones that try to show that moral judgments may be explained by analogy with Chomsky’s grammatically judgments (see e.g. [39] and [30]).

(b) Secondly the research by Haidt and his colleagues devoted specifically to the correlation between moral intuitions and low/high socio-economic status shows that a high level of education, wide variety of contacts, as well as having a large amount and high quality of experiences and knowledge (features which typically go with a high socio-economic status) strongly lead the individuals’ moral intuitions to focus mostly or exclusively on this ‘hard core’ rather than on other aspects. This suggests that – when the individuals’ level of education (in a wide sense) increases – they ‘learn’ to distinguish a moral and a conventional domain according to the criteria put forward by Turiel. That is to say that they learn to identify properly moral violations related exclusively to issues of welfare, justice, and rights and to distinguish them from other kinds of violations, concerning disgusting or disrespectful, but harmless actions. (see [29]).

In this sense, one could say that Haidt’s studies show indirectly that a higher level of education (in a wide sense) acts as a ‘natural antibody’ against the tendency exhibited by people of a low socio-economic status to extend

the domain of moral violations to harmless actions, which are nevertheless considered as disgusting or disrespectful. So, we could suggest that a high level of education (i.e. a high socio-economic status) rescues the moral positions of people from the dominance of the principles they take from their culture and establishes a transcultural connection or unity around the principle that morality has to do with harm and fairness only, while disrespectful or disgusting actions as well as contraventions of religious prescriptions or 'good manners' cannot be considered moral violations.

The fact that this principle seems to be present in all social groups corroborates the idea that questions related to harm-fairness are a necessary part of the human sense of morality. Still, the idea of morality developed by many cultures incorporates also other aspects like in-group/loyalty, authority/respect and purity/sanctity, whose respect is useful for the survival of the culture and of the social order within its borders. Nevertheless, these aspects are not a proper expression of an authentic moral sense, as this is represented by common sense.

As has been pointed out e.g. by social psychologists, loyalty to - and more generally favoritism towards - the in-group is typical in all intergroup relationships. Still, this phenomenon is not necessarily positive from a moral point of view. On the contrary, it sometimes leads to morally negative consequences: in fact, acting unfairly or violently toward people that do not belong to the in-group or, in the worst case scenario, racism towards or infra-humanizing the out-group are possible consequences of in-group loyalty. The same point also applies to respect for authority. To respect authority isn't always or necessarily morally positive: when the orders issued by an authority are not morally admissible, the truly moral reaction is refusing to follow them. And this presupposes once again that the capacity to think in a critical and autonomous way is required in order to act morally and that neither respect for authority nor loyalty towards the in-group are *per se* moral attitudes.

The case of purity/sanctity is similar, although not identical in the sense that it does not describe a social phenomenon, but rather virtues defined by a religious tradition. Each religious tradition has developed its own ideals of purity and sanctity and imposes these on its followers not only through teachings, but also by violently forcing people to embrace them. Historically, respect for ideals like those of purity and sanctity often required the personal sacrifice of sexuality, health or even life. Such sacrifices are perceived as morally positive only by the followers of the religion that imposes them and often only by the most fanatic ones, while persons external to the group or even members of the group who have distanced themselves from the most strict - one could say: inhumane - aspects of their religion consider their imposition as morally impermissible.

These characteristics of in-group/loyalty, authority/respect and purity/sanctity are incompatible with the idea that the higher expression of morality is the capacity of a person to resist blindly following the rules of his/her group or culture, but to autonomously evaluate whether they are good or acceptable

from a moral point of view. According to the principles of in-group/loyalty and authority/respect, for example, the dissident never acts in a morally right manner, since he/she is not loyal to his/her in-group and disrespects the authority within his/her group, i.e. the laws and the rules of his/her group. As far as purity/sanctity is concerned, not only it is highly culturally variable what specifically should be considered pure or saintly, but in many cultures or subgroups these notions have been completely dismissed. As with the case of in-group/loyalty and authority/respect it is easy to imagine the case of a dissident who fights against the idea of purity and sanctity defended by his/her group, whose actions are nevertheless considered unanimously highly moral by all persons external to the group, and even by his/her group later in time.

(a) These remarks give further support to the idea of a lowest common denominator of moral sense focused on harm-fairness-based violations, i.e. *they bear out the thesis of a trans-group moral similarity, in contrast to Heidt's position that morality is highly culturally dependent*. On the contrary, *moral-ity seems to be intrinsically connected with the capacity to think critically and autonomously* from culturally transmitted principles. (b) Furthermore, our analysis points out that a higher level of education modifies people moral thinking and strongly influences the individuals' moral intuitions to focus mostly or exclusively on this common denominator rather than on other aspects connected mainly with cultural beliefs about how one should behave.

If we admit that a primary consequence of a higher level of education is that people's concepts become broader – in particular, that people learn that some beliefs are highly dependent on specific cultures and religions (i.e. on the 'enculturation' each of us is a victim of) – and that people learn and develop the habit of reasoning in a more explicit and critical manner, than these data confirm a continuist interpretation of moral judgment. In fact, they suggest that – even though people often have moral intuitions that comply with the principles of their culture – these are neither the only moral judgments people can reach nor the best possible moral judgments humans can aim for. When we are able to reason slowly and explicitly and to rely on better and more broadly defined concepts, our moral responses can potentially become more 'moral' in the sense that they can overcome cultural factors and limitations and give rise to a transculturally accepted moral stance that best expresses our common moral sense.

## 6 Conclusion

On the basis of our discussion we would like to conclude that a continuist interpretation of intuitionism and rationalism seems to be more plausible than a discontinuist one, since an inferentially and logically based information process offers – at least potentially – better theoretical instruments to explain

moral judgments in the various and flexible forms they assume in different contexts and situations.

Within the continuist approach we subscribe to neither slow nor fast processing is assumed to be immune from errors; however, since intuitions work on the basis of information that is limited and filtered, they are not our best shot in order to form a truly moral stance: the moral stances formed on the basis of intuitions run the risk of not being moral at all. On the other hand, the continuist interpretation must not be mistaken for an abstract ideal of moral reasoning, according to which reasoning may rely on unlimited resources and information. Indeed, reasoning is meant as information processing working on logical and inferential relations that refer to the semantic properties of the information processed. So defined, reasoning turns out to be bounded – i.e. on the concepts we have. Furthermore, according to the position we present here, to reason may not be the first and spontaneous reaction humans have when they face a new situation. In fact, it is often when people meet an impasse, a contradiction or a novel or surprising situation that the natural biases of the cognitive system may be overcome, abandoning the use of heuristics or stereotypes and incorporating more information.

The continuist interpretation challenges the hypothesis of moral intuitionism according to which moral intuitions are rigidly driven by culturally learned principles. We argue that the example of moral dissidence shows clearly that humans are equipped with a moral sense which can be independent from the moral principles recognized by one's own culture, group or subgroup. Furthermore, the analysis we propose suggests that since intuitions are the product of our "enculturation", they are heavily undermined by cultural biases and are, as such, not the best means we have to come to a truly moral judgment. One way to allow culture to play a positive rather than a restraining role is through education. Since a higher level of education contributes both to enriching our knowledge, thereby broadening our horizons cross-culturally, and developing the habit of reasoning in a more explicit and critical manner, we suggest that a higher level of education *potentially* (i.e. given the proper motivation) increase our capacity to reason and to reach judgments free of prejudices and therefore better formulate a moral point of view. In this, we agree with Petty and Wegener [46] when they state that people with low capacities for high elaboration and/or low motivation tend to be easy victims of biasing effects caused by their prejudices, while conscious elaborative processes help to get rid of these biases.

## References

1. Allport, G.W.: *The Nature of Prejudice*. Addison-Wesley, Reading (1954)
2. Banaji, M.R., Hardin, C.D.: Automatic Stereotyping. *Psychological Science* 7, 136–141 (1996)



3. Bodenhausen, G.V., Macrae, C.N.: Stereotype Activation and Inhibition. In: Wyer, R. (ed.) *Stereotype Activation and Inhibition*, Erlbaum, Mahwah (1998)
4. Dellantonio, S., Job, R.: Morality According to a Cognitive Interpretation: A Semantic Model for Moral Behavior. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) *Model-Based Reasoning in Science and Technology*. SCI, vol. 314, pp. 495–517. Springer, Heidelberg (2010)
5. Devine, P.G.: Stereotypes and Prejudice: Their Automatic and Controlled Components. *Psychological Science* 56, 5–18 (1989)
6. Devine, P.G., Monteith, M.J.: Automaticity and Control in Stereotyping. In: Chaiken, S., Trope, Y. (eds.) *Dual-Process Theories in Social Psychology*. Guilford, New York (1999)
7. Evans, J.S.B.T.: In Two Minds. *Dual-Process Accounts of Reasoning*. *Trends in Cognitive Sciences* 7, 454–459 (2003)
8. Evans, J.S.B.T.: How Many Dual-Process Theories Do We Need? One, Two, or Many? In: Evans, J.S.B.T., Frankish, K. (eds.) *Two Minds. Dual Processes and Beyond*, Oxford University Press, Oxford (2009)
9. Evans, J.S.B.T., Frankish, K.: In *Two Minds: Dual Process and Beyond*. Oxford University Press, Oxford (2009)
10. Evans, J.S.B.T., Over, D.E.: *Rationality and Reasoning*. Psychology Press, Hove (1996)
11. Fiske, S.T.: Examining the Role of Intent. Toward Understanding its Role in Stereotyping and Prejudice. In: Uleman, J.S., Bargh, J.A. (eds.) *Unintended Thought*. Guilford, New York (1989)
12. Fiske, S.T., Taylor, S.E.: *Social Cognition*. McGraw-Hill, New York (1991)
13. Fodor, J.A.: *The Modularity of Mind. An Essay on Faculty Psychology*. MIT Press, Cambridge (1983)
14. Fodor, J.A.: *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*. MIT Press, Cambridge (1987)
15. Fodor, J.A.: *Concepts. Where the Cognitive Science Went Wrong*. Clarendon Press, Oxford (1998)
16. Frankish, K.: Systems and Levels: Dual-system Theories and the Personal-Subpersonal Distinction. In: Evans, B.T., Frankish, K. (eds.) *Two Minds. Dual Processes and Beyond*. Oxford University Press, Oxford (2009)
17. Gigerenzer, G.: *Gut Feelings: The Intelligence of the Unconscious*. Penguin, London (2007)
18. Gigerenzer, G.: Moral Intuition = Fast and Frugal Heuristics? In: Sinnott-Armstrong, W. (ed.) *Moral Psychology. The Cognitive Science of Morality: Intuition and Diversity*, vol. 2. MIT Press, Cambridge (2008)
19. Gigerenzer, G., Todd, P.M.: *The ABC Research Group: Simple Heuristics That Makes Us Smart*. Oxford University Press, Oxford (1999)
20. Goldstein, D.G., Gigerenzer, G.: Fast and Frugal Forecasting. *Journal of Forecasting* 25, 760–772 (2009)
21. Habermas, J.: *The Theory of Communicative Action: Reason and the Rationalization of Society*, vol. 1. Beacon Press, Boston (1985)
22. Haidt, J.: The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108, 814–834 (2001)
23. Haidt, J.: The Moral Emotions. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (eds.) *Handbook of Affective Sciences*. Oxford University Press, Oxford (2003)

24. Haidt, J., Bjorklund, F.: Social Intuitionists Answer Six Questions About Moral Psychology. In: Sinnott-Armstrong, W. (ed.) *Moral Psychology. The Cognitive Science of Morality: Intuition and Diversity*, vol. 2. MIT Press, Cambridge (2008)
25. Haidt, J., Graham, J.: When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize. *Social Justice Research* 20, 98–116 (2007)
26. Haidt, J., Graham, J.: Planet of the Durkheimians, Where Community, Authority, and Sacredness are Foundations of Morality. In: Jost, J., Kay, H.T.A.C. (eds.) *Social and Psychological Bases of Ideology and System Justification*, Oxford University Press, Oxford (2009)
27. Haidt, J., Joseph, C.: Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus* 133, 55–66 (2004)
28. Haidt, J., Joseph, C.: The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Cultural-Specific Virtues, and Perhaps Even Modules. In: Carruthers, P., Laurence, S., Stich, S. (eds.) *The Innate Mind. Foundations and the Future*, vol. 3, Oxford University Press, Oxford (2007)
29. Haidt, J., Koller, S., Diaz, M.: Affect, Culture and Morality, or it is Wrong to Eat Your Dog? *Journal of Personality and Social Psychology* 65, 612–628 (1993)
30. Hauser, M.D.: *Moral Minds. How Nature Designed Our Moral Sense of Right and Wrong*. Collins Publisher, New York (2006)
31. Kahneman, D., Slovic, P., Tversky, A.: *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge (1982)
32. Kahneman, D., Tversky, A.: *Choices, Values, and Frames*. Cambridge University Press, Cambridge (1982)
33. Kant, I.: *The Critique of Practical Reason*. Cambridge University Press, Cambridge (1788-1977)
34. van Knippenberg, D., van Knippenberg, A.: Social Categorization, Focus of Attention and Judgments of Group Opinions. *British Journal of Social Psychology* 33, 477–489 (1994)
35. Kohlberg, L.: Stage and Sequence. The Cognitive-Developmental Approach to Socialization. In: Goslin, D.A. (ed.) *Handbook of Socialization Theory and Research*, Rand McNally, Chicago (1969)
36. Macrae, C.N., Hewstone, M., Griffiths, R.J.: Processing Load and Memory for Stereotype-Based Information? *European Journal of Social Psychology* 23, 77–87 (1993)
37. Macrae, C.N., Milne, A.B., Hewstone, Bodenhausen, G.V.: Stereotypes as Energy-Saving Devices: A Peek Inside the Cognitive Toolbox. *Journal of Personality and Social Psychology* 66, 37–47 (1994)
38. Magnani, L.: *Abduction, Reason, and Science. Processes of Discovery and Explanation*. Kluwer Academic, New York (2001)
39. Mikheil, J.: *Rawls Linguistic Analogy*. Ph.D. Thesis Cornell University Press, New York (2001)
40. Nerveaz, D.: The Social Intuitionist Model: Some Counter-Intuitions. In: Sinnott-Armstrong, W. (ed.) *Moral Psychology. The Evolution of Morality: Adaptations and Innateness*, vol. 1. MIT Press, Cambridge (2008)
41. Nisbett, R.E.: *The Geography of Thought. How Asians and Westerns Think Differently... and Why*. Free Press, New York (2003)

42. Nisbett, R.E., Cohen, D.: *Culture of Honor: The Psychology of Violence in the South*. Westview Press, Boulder (1996)
43. Pastore, L.: *Intuition*. In: Sandkühler, H.J. (ed.) *Enzyklopädie Philosophie*. Felix Meiner Verlag, Hamburg (2010)
44. Peirce, C.S.: *Collected Papers*. 1903 *Harvard Lectures on Pragmatism*. In: Hartshorne, C., Weiss, P.(ed.) vol. 5 (vols. I-VI). Harvard University Press, Cambridge (1931-1958)
45. Petty, R.E., Cacioppo, J.T.: *The Elaboration Likelihood Model of Persuasion*. In: Berkowitz, L. (ed.) *Advances in Experimental Social Psychology*, vol. 19. Academic Press, Orlando (1986)
46. Petty, R.E., Wegener, D.T.: *The Elaboration Likelihood Model: Current Status and Controversies*. In: Chaiken, S., Trope, Y. (eds.) *Dual Process Theories in Social Psychology*. Guilford Press, New York (1999)
47. Piaget, J.: *The Moral Judgment of the Child*. Free Press, New York (1932-1965)
48. Sloman, S.A.: *The Empirical Case for Two Systems of Reasoning*. *Psychological Bulletin* 119, 3–22 (1996)
49. Smetana, J.: *Understanding of Social Rules*. In: Bennett, M. (ed.) *The Development of Social Cognition: The Child as Psychologist*, Guilford Press, New York (1993)
50. Smith, E.E.: *Concepts and Categorization*. In: Smith, E.E., Osherson, N.D. (eds.) *Thinking. An Invitation to Cognitive Science*, vol. 3. MIT Press, Cambridge (1995)
51. Smith, E.R., DeCoster, J.: *Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems*. *Personality and Social Psychology Review* 4, 108–131 (2000)
52. Stangor, C.: *The Study of Stereotyping, Prejudice, and Discrimination Within Social Psychology. A Quick History of Theory and Research*. In: Nelson, T.D. (ed.) *Handbook of Prejudice, Stereotyping and Discrimination*. Psychology Press, New York (2009)
53. Stanovich, K.E.: *Who Is Rational? Studies of Individual Differences in Reasoning*. Lawrence Erlbaum Associates, Mahwah (2009)
54. Tisak, M.: *Domains of Social Reasoning and Beyond*. In: Vasta, R. (ed.) *Annals of Child Development*, vol. 11. Jessica Kingsley, London (1995)
55. Turiel, E.: *The Culture of Morality. Social Development, Context and Conflict*. Cambridge University Press, Cambridge (2002)
56. Turiel, E., Killen, M., Helwig, C.: *Morality: Its Structure, Functions, and Varieties*. In: Kagan, J., Lamb, S. (eds.) *The Emergence of Morality in Young Children*. University of Chicago Press, Chicago (1987)
57. Uleman, J.S., Bargh, J.A.: *Unintended Thought*. Guilford, New York (1989)
58. Walton, D.: *Abductive Reasoning*. The University of Alabama Press, Tuscaloosa (2004)

# Evolutionary Tolerance

Luís Moniz Pereira

**Abstract.** The mechanisms of emergence and evolution of cooperation — in populations of abstract individuals with diverse behavioral strategies in co-presence — have been undergoing mathematical study via Evolutionary Game Theory, inspired in part on Evolutionary Psychology. Their systematic study resorts as well to implementation and simulation techniques in parallel computers, thus enabling the study of aforesaid mechanisms under a variety of conditions, parameters, and alternative virtual games. The theoretical and experimental results have continually been surprising, rewarding and promising.

Recently, in our own work we have initiated the introduction, in such groups of individuals, of cognitive abilities inspired on techniques and theories of Artificial Intelligence, namely those pertaining to Intention Recognition, encompassing the modeling and implementation of a tolerance/intolerance to errors in others — whether deliberate or not — and tolerance/intolerance to possible communication noise. As a result, both the emergence and stability of cooperation, in said groups of distinct abstract individuals, become reinforced comparatively to the absence of such cognitive abilities.

The present paper aims to sensitize the reader to these Evolutionary Game Theory based studies and issues, which are accruing in importance for the modeling of minds with machines. And to draw attention to our own newly published results, for the first time introducing the use of Intention Recognition in this context, with impact on mutual tolerance.

**Keywords:** Evolutionary Game Theory, Evolutionary Psychology, Intention Recognition, Tolerance.

## 1 Evolution and the Brain

Darwin's hypothesis about the biological evolution through natural selection was one of the most revolutionaries ones in the history of science. Since the

---

Luís Moniz Pereira  
Centro de Inteligência Artificial — CENTRIA  
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa  
2829-516 Caparica, Portugal  
e-mail: lmp@fct.unl.pt

publication of the *On the Origin of Species* in 1859, and until today, a long history of attempts at applying the evolutionary concepts to the understanding of human and social behavior has occurred. Some of these, more polemic, attempts, had disastrous political interpretations and applications (such as, for example, the defense of white supremacy by Nazism), and several intellectual groups developed antibodies against the widening of the scope of evolutionary concepts. Other attempts revealed more fruitful. One such example is the attachment theory developed by Bowlby (1971), extrapolated to the evolutionary processes by Kirkpatrick (2005), with solid empiric evidence and a vast explanatory power. Bowlby considered attachment as an organized set of behaviors, which evolved through the mechanisms of natural selection, to solve a recurrent adaptive problem: the need for protection that immature members have, as much amongst humans as amongst other species.

The *Homo sapiens sapiens* emergence is consensually situated within the Superior Paleolithic, around 45 thousand years ago — by the time when language, sedentary and gregarious behavior had fully developed. One assumes that, in terms of cerebral morphology, we would have been by then essentially the equal of what we are nowadays. From about 40 years ago, the discipline of Evolutionary Psychology has been developed founded on the application of evolutionary concepts to the understanding of the psychological mechanisms that underlie human conduct. It provides a way of thinking about evolution when the advantage of a given behavior depends on some other individual's behavior, or on a group's. According to its followers, there are several human behaviors that are better understood if we rebuild the way how natural selection acted in the past, and lead to the emergence of *Homo sapiens sapiens*.

The first bipedal primates established the separation between the human species and other apes. In order to encompass the capacities of the human brain one needs to understand exactly which problems our primate ancestors were trying to solve, and what lead them to develop such an extraordinarily intricate brain. We cannot look at the modern human brain, and its capacity for creating science, as if the millions of years of evolution that shaped it till its present form had not happened. Amongst the eventual problems there are certainly those of *status*, territoriality, mating, gregarious behavior, altruism versus opportunism, the construction of artifacts, and the mapping of the outside world. The brain of the *Homo sapiens sapiens*, considered anatomically indistinguishable from our current brain, is approximately 100 to 200 thousand years old, with oral language appearing less than 100 thousand years ago. The Superior Paleolithic began around 45 thousand years ago, and lasted until the Neolithic, about 10 thousand years before the current era (C.E.) — during this period language fully matured. However, since the beginning of the Superior Paleolithic, the cultural evolution rhythm has drastically accelerated. According to population genetics theory, the majority of changes happened too quickly to be accompanied by genetic evolution.

The same way a psychiatrist or a psychoanalyst has to look at a patient's history in order to better understand him in the present, it is also important to look at our specie's past in order to grasp our modern peculiarities.

Evolutionary Psychology began with sociobiology and the study of insect societies. The quest was to discover the why and how these animals are gregarious. Research was developed in the early 1960's by William D. Hamilton (1926-2000), Robert L. Trivers (b. 1943), and later by Edward O. Wilson (b. 1929). That research was carried out mathematically, first in terms of game theory and computer simulations, and then continued with contributions from other disciplines.

Highly altruistic, the social insects enjoy the so-called haplo-diploidism (instead of our own diploidism), which makes siblings share more genes than usual. In the females of those insects (bees, ants, termites) half the DNA is an exact copy of the father's haploid DNA, and the other half is from the diploid mother — they thus share, on average, 3/4 of their genes. The fact they share more genes endows them with a greater predisposition to sacrifice themselves for their siblings. It is this genetic mechanism that induces a greater social cohesion, and a greater altruism, because it is genes that survive (the «selfish genes») and not the vehicles of the genes, the living beings that transport them like dispensable packaging (Hölldobler & Wilson, 2009).

The problem of selection is particularly important concerning the consideration of individual and group components. Beyond a simple survival of the individual — or family — there is the survival of a larger group which, in a gregarious species like ours, is of extreme importance.

And the problem in explaining cooperation by evolution is this: “By which mechanisms are we the product of gregarious evolution, in that gregarious behavior benefits everyone? How altruistic and socially cooperative are we, or, being altruistic, won't we be fooled by others, the opportunists, focused singly on individual selection?” The evolution of any collective species clashes against this problem of balancing altruism with egotism. It is a strong theme in Evolutionary Psychology, and one to which we can employ computers to perform long and repetitive simulations of joint evolution of behavioral strategies in co-presence, via typical mathematical games' implementation, mixing competition and cooperation situations, and the combinatorics of strategies.

Evolutionary Psychology is not so much a scientific discipline but more of a meta-theoretical reference framework, of assumed presuppositions shared by researchers working in the field. As a starting point it, considers human brains are the product of the evolutionary process and that this fact cannot be ignored when trying to understand the workings of the mind and of behavior.

In this manner, Evolutionary Psychology is becoming a success example under the scope of ongoing scientific unity, resulting from a profound and explanatory combination of Psychology, Anthropology, Biology, Linguistics, Neurosciences, Game Theory, and Artificial Intelligence (Laland & Brown, 2002; Buss 2005; Dunbar & Barrett, 2007; Gangestad & Simpson, 2007; Platek et al., 2007; Skyrms 1996, 2004, 2010). It has been dedicated to the study of the brain and of behavior from an evolutionary perspective, having given rise to extremely relevant contributions. And it has been backed up, and influenced, by Anthropological Archeology in its empirical study of the cultural evolution of mankind (Shennan, 2002). Along this line of development, Evolutionary Psychology has been

revealing itself as a paradigm of analysis which is very rich and useful for the understanding of universal sexual differences in the strategies used by men and women when choosing a mate — men tend to seek several young women, whereas women tend to select a unique partner with characteristics associated with power; men tend to be more violent than women; etc. In the study of the workings of the brain through their archeological traces, both theoretical as well as field archeologists (Mithen, 1996), are bringing about historical and pre-historical evidence that our ancestors began with a generic intelligence, such as we find nowadays in apes. There has been intense and wide-scope discussions on the problem of intelligence being of a generic functionality, or being better understood instead through division into components or modules of specific abilities. When Evolutionary Psychology first appeared, it developed a line of work, which Chomsky had started, that insisted in the existence of innate and specialized areas of the brain; it was generally accepted that there is an abundance of specific modules for a diversity of cerebral functions. In the beginning, the opinions of David Buss, Leda Cosmides, Steven Pinker, John Tooby (Buss, 2005), pointed to the scenario where all cerebral functions had their own specific modules — for language, for mating, for religion, etc.

Meanwhile, through historical record, archeologists showed that the human species went from a first stage of generic intelligence to a second stage comprising three big specialized modules: one dedicated to natural history and the rudiments of physics (knowledge of Nature); another for the knowledge and fabrication of instruments; and a third for the cultural artifacts, i.e., the rules of living in society and the politics of coexistence. These three specialized intelligence types were separated. However, at a more recent stage — corresponding to *Homo sapiens*, with the appearance of spoken language — it became necessary to have an umbrella module able to articulate all the other three. And the question arises: How do all these different specialized modules connect, and how do people communicate? The need to find an answer to this problem gave birth to the idea of an encompassing and overlaying module, a more sophisticated form of generic intelligence, the cognitive glue that binds together the specialized modules and allows them to communicate and cooperate.

From our point of view, logic, in a broad sense, provides that encompassing and overarching general conceptual cupola which, as a generic module, allows the fluid articulation of the more specific modules. There is an obvious human ability to understand logic reasoning, and such ability must have developed during the evolution of the brain. The computers we create share the similar counterpart ability to execute any program (Pereira, 2009).

## **2 Evolutionary Psychology: Genes and Memes**

The main notion, with which we must begin, is to understand that there are two Darwinian mechanisms in co-evolution in humans. By Darwinian we mean the great paradigm of emergence that results from mutations, selection and reproduction, that brought life to Earth up till today and, in particular, gave rise to human beings as a species.

Life began on Earth about 3,8 billion years ago with bacteria; only 2 billion years after that, the first unicellular organisms with a nucleus — the eukaryotic cells — appeared. Their components got together thanks to collaboration amongst bacteria and, throughout evolution, ever increasingly complex organisms emerged from simpler ones, with millions of cells binding together to form tissues, different tissues cooperating to build organs, and these intertwining to form systems. Amazing similarities can be found in the way biological complexity increases and the way human societies evolve. The importance of cooperation is evident in both biological and social domains of complexity (Damasio, 2010).

There are two reproductive systems in humans: the sexual reproduction one, in which the replication unit is the gene; and the mental reproduction one. Some authors in the Evolutionary Psychology field defined the notion of «meme», as a mental counterpart of the gene. The meme is the mental replication unit, dual to the gene, and its reproductive system is the brain. Memes get together in groups, patterns or «memplexes», in a way similar to the union of genes when they form chromosomes and sequences. Memes are characteristic of ideologies, religions and common sense ideas. Certain memes only work well together, mutually reinforcing one another, and others do not, in such a way that certain correction devices must be triggered into action. Mechanisms of tolerance/intolerance, which we further detail in the sequel, can also be triggered, working both at individual and group levels.

The two Darwinian mechanisms in co-evolution are thus the genetic and the memetic (Dennett, 1995). There is a genetic reproductive system and, on top of it, Nature — through evolution — created a second one, which we employ in pedagogy. We reproduce ideas: normally the good ones are propagated and multiply, being selected for, in detriment of worse ones — although nothing and no one guarantees such selection skewing. Genes persist because they reproduce themselves, while memes comprise a parallel reproduction unit associated with the mind — the brain being its reproductive organ. What we do, in schools and universities, is to reproduce knowledge. Educational systems consist in a means to «infect» students with memes, ideas proven capable enough to reproduce and persist, while others that could not survive were discarded in the process. Of course, there are many variants of educational systems, for instance the madrasahs.

When people interact they communicate ideas, and the infectiously good tend to reproduce. As aforementioned, there are groups of ideas, belief sets, that reproduce together. The memes that are part of such clusters — like genes in chromosomes — are in competition/cooperation amongst themselves, and also with the pool of genes. These exist because they are part of a reproductive system necessary for attaining local adaptations more quickly, knowing that genes, concerning the temporal scale of the meme-carrying individuals, take too much time to reproduce. In this way, the meme rich individual phenotype benefits from another chance to improve the conditions to replicate its genotype. This leads directly to meme-gene co-evolution.

However, memes could not spread if it weren't for the valuable biological tendency individuals have to imitate, something the brain is neurologically capable of, namely via mirror-neurons (Rizzolatti & Sinigaglia, 2007). There are



very good reasons for imitation to have been favored by conventional natural selection acting on the genes. Individuals genetically more predisposed to imitate can take advantage from a shorter path to learning new skills, that others might have taken longer to build. Consequently, the brain and the mind that goes with it are the result of a profound symbiosis, a genetic product influenced by the memetic reproduction mechanism. With this faster adaptation system we arrived at a point where we can predict our own necessary memetic mutations, as preventive measures needed to prepare ourselves for the future, by anticipating it. As a result, we imagine the future — we create hypothetical scenarios, evaluate possible outcomes — and choose to strive towards some of them, calling it «free will».

As a consequence of the existence of the memetic system, beyond single genetic sexual reproductive success, there arise some important issues regarding social interaction. As communal beings we need to develop a *status* in order to be respected, copied from or obeyed. We have to worry about territorial expansion and its defense if we want to possess the resources necessary to have offspring and, what is more, if we desire our offspring to have offspring of their own. We need to take part in contractual agreements with whomever shares our social and cultural ecology. And there exists also the important requisite of opportunity for personal expression. If we do not express ourselves, no one will copy even our most precious memes, let alone our scientific theories built from memplexes.

With this perspective, in spite of a spatial and temporal distance, scientific thinking emerges from distributed personal interaction, and never as an isolated act. This interaction has to be built from the ground up from several confluences, or by teams, as is the case in science. Indeed, knowledge is not constructed in an autonomous way; rather, it is weaved by networks of people. In science it is important to work as a team, and science itself comes institutionalized and organized with its own methodologies. It takes place in particular environments, as is the case of educational ones, where memetic proliferation is mechanized.

### 3 The Logic of Games

Game theory was first developed in the 1940's, and the first work on the subject was *Theory of Games and Economic Behavior* by the mathematician John von Neumann (1903-1957) and the economist Oskar Morgenstern (1902-1977), (Neumann & Morgenstern, 1944). At the time it was directed at the economy, but it was subsequently applied to the Cold War, as the outcome of issues raised by the use of the atomic bomb and the subtle means of bluffing. When some such situation gets complicated, there is need to resort to sophisticated mathematical tools — and computer simulations — to deal with equations that cannot otherwise be solved.

The games theme is as complex as it is interesting and filled with diverse niches. We already addressed genes, memes, their combinations and evolution, questions related to survival and winners. We have already mentioned the combinatorial evolution of strategies, and mutations of those strategies according to diverse conditions, that can both be other game partners or the game board's own circumstances. The notion of game includes uncertainty, and whenever there

is uncertainty there has to be some strategy, the moves one makes with given probability. When there is co-presence of evolving strategies from several partners, along with the idea of payoff, we are dealing with the notion of evolutionary game, which can be examined in an abstract and mathematical manner.

The same way we have genetic strategies for reproduction, all of our lives are filled with cultural, or civilizing, strategies. And, in a general way, we can see our species through these lenses still without undervaluing the remaining perspectives, equally important.

There are zero sum games and non-zero sum games. The zero sum ones are those that, by their rules, some players win, some players lose. In Nature's evolution, conditions are those of non-zero sum — all can win or all can lose. Robert Wright (2001) analyses the evolution of culture and civilization with the underlying idea that, in Nature, non-zero sum games are possible, wherefore a general gain may be obtained, leading to illuminated altruism.

Sometimes, co-present strategies tend to achieve a tactical equilibrium. Take the hunter/prey relationship: neither the hunter wants to fully exterminate the prey, nor the latter can multiply indefinitely because that would exhaust the environment's resources. Some of these studies are used by Economics to understand what might be the overall result from the sum of interactions amongst the several game partners.

It is relevant to take into consideration if the game takes place only once with a given partner, or whether the same partner may be encountered on other occasions; how much memory does one have of playing with that partner; and if the possibility of refusing a partner is allowed. Let us take a more detailed look at each of these situations in turn. We begin with the famous prisoner's dilemma, typical of the paradox of altruism. There are two prisoners, A and B, with charges on them. Either of them can denounce the other, confess, or remain silent.

	Prisoner <b>B</b> – silence	Prisoner <b>B</b> – confession
Prisoner <b>A</b> – silence	<b>6 years in jail for each</b>	<b>A = 10 years in jail</b> <b>B = 2 years in jail</b>
Prisoner <b>A</b> – confession	<b>A = 2 years in jail</b> <b>B = 10 years in jail</b>	<b>8 years in jail for each</b>

Let the above be a 2x2 payoff matrix where the lines correspond to the behavior of A (remain silent or confess) and the columns correspond to the behavior of B (remain silent or confess). At the intersection of B's «confess» column with A's «confess» row, both receive a jail sentence of 8 years. If A confesses and B does not, A will only get a 2-year sentence, whereas B gets 10 years, and vice-versa. There is an incentive for any of them to confess in order to

reduce their own jail sentence. This way, it would eventually be advantageous for them not to remain silent. If one of them defects by confessing, but not the other, he will only stay in jail for 2 years whereas the other will be there for 10 years. But if both confess they will be sentenced to 8 years each. The temptation to confess is great, but so is the inherent risk, because, after all, they would mutually benefit from remaining silent, getting a 6-year sentence each in that case.

The prisoners know the rules of the game, they just do not know how the other player will act. It is advantageous for them to remain silent, but they do not know if the other one will confess. As long as one of them confesses, the silent other will be sentenced to 10 years in jail. A dilemma thus arises: it is good to remain silent, but there is the risk the other one will defect; and the one who does it faster will take the greater gain. In the worst case scenario, both get an 8-year sentence — nobody will take the risk. This is a classic game, one where both players have the tendency to confess — and not benefit from what could be a mutual advantage, but one they cannot assuredly profit from. Firstly, because they do not have the opportunity to talk; secondly, because even if they did, they would still be in risk of being betrayed by the other. They have no joint solution in the sense that A and B could ever choose what is best for both, where there would be an assured increased advantage for the two.

All turns more complicated when one imagines A and B playing this game many times in succession, taking into account their experience of previous mutual behavior in their past. In this case they can go on building mutual trust or distrust. If one betrayed the other once, the betrayed one's reaction will be vengeance, or simply intolerance, in some future opportunity. Let us now visualize a situation with multiple players and ask ourselves which will be, along time, the best of all possible strategies — by running a computer simulation. Of course one thing is to presuppose any one strategy can always match with any other, which is the base assumption, and then to move on to a situation where one wants to match only with certain players. Through these more realistic situations one begins to develop a game theory where social structure is included inside it.

In the early 1980's, Robert Axelrod (1984) launched a worldwide competition taking place in computers, by setting up the following game: Each participant programs, on the computer, the strategy he intends for his player-program in the Prisoner's Dilemma game; thus there is a population of player-programs written by the participants, each with its own specific strategy. At each successive step of the game running on the computer, each player is matched with another player, and each either plays «cooperate» or «defect» according to their respective strategies. If they both cooperate they both win, according to the payoff-matrix settled for the game; if only one defects it will benefit, but the betrayed player will have the chance to reformulate its future behavior. Which will be the best strategy to take when this game is played over and over for thousands of iterations in co-presence with other strategies? The best strategy actually depends on which other strategies are present but it was shown — and the experiment was repeated with a different collection of strategies — that the then best option was the so-called «Tit-For-Tat (TFT)». This strategy consists in beginning by cooperating and, from that point onwards, repeating whatever the adversary did just before — in plain English,

"what comes around goes around". If someone betrays me, on principle I will also betray in the next round because I am imitating him. If from the start both cooperate having this strategy, then they will receive back successive cooperation. Amongst all possible choices — cooperate three times, betray twice, etc. — it was shown that the TFT option is, par excellence, a winning strategy which tends to invade a population of strategies. «Invasion» because in presence of someone with a winning strategy, one will imitate it and, if we all play TFT we all win the most. All start by cooperating and keep on doing it, and thus always keep on winning whatever payoff comes out of cooperation. There is only one circumstance where TFT does not invade the population: when all other players are betrayers and only one cooperates — because then there is no one to cooperate with. However, as long as there are two TFT players in the population, they will win whenever they meet and thus start invading the population on being imitated.

Instead of imitating those who win the most, one can alternatively let them reproduce the most, that is, make more copies of them, proportionately, and keep a fixed size for the population, by random elimination. This option can be had because those who lose more easily are eliminated by virtue of their reduced number of copies, and because only those who win more than some threshold are allowed to reproduce. The intent of this interpretation is that, throughout the game, we want to take over resources and occupy vital space for the population. Winning means having more energy to reproduce, while losing means not being able to persist with one's genetic/memetic continuity.

The winning strategies invade the population and, since they self-support, they are labeled "evolutionary stable". Things tend to complicate, naturally. From the moment we introduce more memory, it is possible to remember how much a given individual betrayed us, and who didn't, evaluate them against our own betrayals, and thence exercise tolerance, or not. Our strategy then is not blindly applied to each individual met, but for doing so we need memory, even if limited. If someone plays «Tit-For-Tat» and, every once in a while, betrays, he can accumulate more benefits, in the sense that the other player may exert a tolerant forgetfulness regarding the harm suffered. Till today, with just memory of the last play, the most successful strategy, superior to TFT because resistant to error or noise, is the «Win Stay, Lose Shift (WSLS)» one. That is, if I won in the last play, I repeat my behavior; if I lost, I change it in my next play; I start by cooperating (Sigmund, 2010).

It is also important for a strategy which aims to be evolutionary stable to be tolerant to the inevitable evolutionary noise or imperfect communication, in such a way as to be able to recover from endless cycles of revenge and counter-revenge. Of course, there will be opportunistic strategies that will try to make an intended betrayal look like noise, and thus gather the benefit of doubt forgiveness. Tolerance needs guardedness. Intention recognition thus becomes important, including how somebody else's intentions are affected by the way others recognize and tolerate our own intentions (Pereira & Anh, 2011, 2012, 2012a). In our recently published and submitted work (Anh & Pereira & Santos, 2011, 2011a, 2012, 2012a), we have developed an intention recognizer strategy (IR), which wins against WSLS and against all others too, and is evolutionary stable even in

the presence of substantial noise. With only about 10% of initial IR players, the strategy invades the population of Prisoner's Dilemma players, or of other classic games — such as the «Stag Hunt» — even in the most disadvantageous conditions of the payoff matrix, that is when the gain of betrayal, and thus the temptation of acting on it, is very large comparative to other payoffs.

Another problem also arises concerning the possibility of reencountering or not the betrayer. If you only encounter that player once, the likelihood of betrayal increases. But if, on the other hand, I know that I will encounter the betrayer several times, the chances of betrayal happening become lower. And I know that kind of memory can be communicated to others — I can tell someone: «you can trust that fellow, take my word», and the other person may believe what I tell him because I have gained some credibility. If I never betrayed anyone, I am a friend — our friends can believe in us — and I can spread that information about the credibility of others, thereby ensuring a certain degree of tolerance towards them.

Whenever there is the choice of playing, or not, with a given partner, the whole game situation changes. One can, for example, explore the evolution of social structure. This means that I begin by clustering players into groups who prefer the same strategies. Scientists like to play with other scientists, lawyers play with lawyers — they have the same psychology, they know what they can expect. It is possible to obtain a larger gain by knowing whom each player can relate with. There is, therefore, the tendency to play with whomever we can establish a trust relationship, and whose thinking strategy type is familiar to us, for our own defense.

This relates both to the memetic game as well as to the genetic game. It is the evolution of the civilization «pool», besides the genetic «pool», that matters. The problem cannot be seen in terms of reproduction of the single individual, but as the reproduction of certain shapes and configurations of genes and memes that make society, as a whole, to benefit from the coexistence of that variety of strategies in co-presence. It is moreover possible to prove that, in certain games, it is the combination of strategies that wins, keeping to an equilibrium amongst them — if one is taken away the whole will be harmed. In general this is the way ecological systems work. It is a combination of strategies of various organisms — some parasitizing others — so that the whole system may live, survive and continue to evolve at the cost of such multiple equilibriums.

The so-called culture of altruism, as long as it is shared by the elements of a group, can be a winning strategy. Under said conditions, as mentioned before, an opportunist can always be born, a parasite, he who says to himself: «I've understood the game and this is what I'll do». The others, meanwhile, find out his scheme and create mechanisms to detect and to not tolerate him. However, each adaptation will make him an ever more sophisticated opportunist — for example, the one who discovers a loophole in a law and takes advantage of it. It is very much necessary to know if one can be detected when preparing a betrayal of the members of the group. To the contrary, if we are about to be detected the guilt feeling grows, and it might eventually reach a point where we confess even before we are caught — another way of getting benefits. The production of a guilt feeling may be seen as a strategy for such a situation.

Wanting to be simply altruistic, even when we expect nothing in return — because exchanges need not be immediate — can be advantageous because, sometimes, it may be good for us to have others just seeing us be kind. If someone sees me being altruistic I will be increasing my credibility. And the best way to gain a good reputation is by being effectively good, or faking to be good. But going down the faking road can lead to a point where we are not aware of being fakes. That is so because I deceive others best if I deceive myself too; otherwise I will be too conscious of my faking and could blow my own cover with a misplaced gesture, and no longer be tolerated in the game. Of course this creates conviviality problems, precisely because he who fakes is convinced that he is not faking, as a result of his consciousness elimination mechanism.

The situation can become more complicated in many ways. There is benefit in being altruistic to the extent that we are betting on the safe side. In a society where altruism is beneficial, showing that you comply with the rules of the game will entitle you to rewards, if the day comes when you need them, as long as there is a measure of intolerance against opportunists. We have social security systems, unemployment benefits, etc., but there will always be that individual who takes undue advantage of them, one way or another; and particularly those who say that such systems are useless, so as to justify not to contribute.

However, there are certain cultures and cultural levels — and I am thinking of two — where the most important is to «give». In New Guinea, every six months, there is a ceremony where people offer pigs, and where whoever offers more pigs becomes, for one season, the overlord. The others feel the obligation to give pigs back, so they raise lots of pigs — which is always very difficult because it is necessary to keep feeding them for six months before offering them — but whoever gives more pigs away obtains greater social recognition. This form of altruism and gift are used as social mechanisms to assess status, to assess who is more important. The same happens with Eskimo, in the potlatch ceremonies — a gift offering and wealth distribution ritual — which is equally a celebration of who gives the most. Yet another example are the scientists — they work to do more research and to publish more results for free, to be more considered and gain social returns. The game consists of giving more, not giving the least to gain the most, because the gain is measured in a different currency.

Another aspect of a player is the notion of honor, most exacerbated in the 19th century, which is the idea of earning a good reputation and knowing how to preserve it, if necessary by means of the courage of strength. When an individual is offended, he takes out a glove and strikes back at his offender slapping him twice — even if the offense is nothing of much importance. What is at stake is to know that the offended always responds. In a sub-organized society — in which altruism is not generalized and in which not everyone pays back — it is necessary to have a signal that defines «I am one of those who pays back, and in the name of my credibility and my honor this is how I behave».

To understand these mechanisms and, opportunistically, profit from these theories, allows whomever knows them to take advantage of the credulity of others.

I have already mentioned that the next step is to have a language system, of commitment by language and, of course, of deceit. One can say «I am all in favor of strategy A» and the other one responds: «I am in favor of strategy B», each one having certain gestures or labels that identify them. Signals are given, which is strategically beneficial because it avoids confusions, betrayals or surprises. Brian Skyrms (Skyrms 1996, 2004, 2010), currently one of the great scholars in this area, shows how systems of signals can arise which, according to him, evolve and are at the origin of languages. Signals correspond to the necessity of determining which kind of player one is, and which secret codes one has access to. If I am with the Freemasons, and the other person says he is too, I will ask him to prove it and he will give me a special handshake. And only those who know it will respond correctly, which allows to identify him as a legitimate member of the group. Of course there are also imitators, and where there are imitators there is also someone that detects and not tolerates them. Bottom line, this “arms race” puts a prize on the evolution of our brains through the games we are involved in. But, I repeat, the strategies are not of one single individual, but of a whole society or group — which can be political, religious, etc. In other words, it is very important to know which partners we are playing with.

This capacity to choose partners raises, a priori, the question of knowing which kind of partner we can rely on. The casting of group identifying signals mentioned above is quite economic, in terms of games; it is more economic than a play, and it can be very rewarding. This way, game theory with signals is born. But when there are signals, there are those who pretend, there are codes that must remain secret.

The choosing of partners thus creates preferential groups, with their unifying and protecting advantages. But the negative side of that is the definition of borders between groups, and the coming into force of group competition for resources, which places and repeats at a new level and scale the problem of altruism/opportunism and that of tolerance/intolerance.

## 4 Strategic Equilibriums

Game theory focuses on the average expected gain. No one predicts the future and one must simulate all possibilities, with a broad sense of what are the respective probabilities of occurrence. A topic of study is the so-called «Nash Equilibrium». This is about a strategic equilibrium point where we disadvantage ourselves if any of us changes our payoff percentage, increasing or decreasing it. Let us illustrate a few phenomena of evolution towards an equilibrium, experimentally observed in lab and field games specially concocted.

Suppose the following game, where someone with a cake says: «I have a cake to split among us», the whole cake is 100% and each of us must write down the percentage he wants. If someone writes down that he wants 80%, and I myself do the same, the total will surpass 100% and in that case nobody gets any part of the cake. If the sum does not surpass 100% each one gets what he wrote down. If an individual writes down 30% and I write 70% we get the whole cake and share it that way — this is an example of a «Nash Equilibrium». If anyone of us increases

his percentage, even if by only 1%, the sum will surpass 100% and we lose all the cake. If we decrease our percentage, we get less than 100% of the cake, which we could have had. This game has an infinite number of solutions, as long as the total adds up to 100% — we achieve the equilibrium and nothing prevents us from playing that way. But if I see somebody eating 70% I will ask for another cake, and the splitting will evolve towards 50/50. Why? The point is that, usually, the strategy occurs within a group. We are many, there are lots of people with cakes and we are all eating them. When I see a couple doing 50/50, and I check that it works, my imitation mechanism will intuitively lead me to copy it. There is an imitation mechanism because those going for 50/50 get the whole cake, get fat and have more offspring with their investment. If they reproduce more, this strategy tends to invade those of others. Those who opt for 30% of the cake, get thinner, will progressively grow weaker, cannot afford the luxury of having offspring — and, in the long run, their strategy will be wiped out. The 50/50 tend to invade the playing field not only by imitation but also by genetic mutation.

In fact, strategies are not fixed. Today I say 30%, tomorrow I will say 35% and, as a living being, I automatically explore around an equilibrium. I try other strategies and the most beneficial persist. The 70/30 split is unstable only if there are others using different strategies, if imitation and crossover are in place. The mutation evolutionary component, the reshuffle and distribute, is essential to attain a «Nash Equilibrium» and to allow evolution towards another solution. Let us return to the cake game, but let us substitute it by a 100 dollar bill that we are splitting in two. Suppose I ask for 30 and the other asks for 70. But let us modify this a bit — there will be a referee deciding who gets the 70 and who gets the 30. If I know the referee and I bribe him I will ask for 99% or even 100%, and he will agree with me. But if, on the contrary, he is fair and I do not know him, it is obvious I will ask for 50%. On the other hand, if the game is not decided by a referee but randomly, say, by a computer generating a random solution, the result would be different. The fact of whether there is or there is not a referee, as well as the kind of referee, will make a big difference — the equality of distribution is now associated with the need for a justice system with judges — and this is how it arises. We will find the 50/50 arise, which our intuition tells us is correct, so much it is branded in our genes. This problem is so general and common that it had to be solved for millions of years, time and again.

There is another game to consider, involving four people. Initially a scientist, in a laboratory experiment, equally distributes a certain amount of money by four people, and these must decide with how much they will contribute to a common fund, which the scientist will then double and use for a new distribution. Suppose I contribute with 10%, the others do the same and the resulting total will be again equally distributed. The problem is evident. The people contribute with equal amounts, the scientist adds a new contribution equal to the sum of the four contributions, thereby duplicating it, and then redistributes it by the four. That is, there is an advantage in contributing to the total because the sum will be double. But I can also get an advantage if I do not contribute to the sum — because others will and, anyway, I will benefit.



These are the explicit rules, and most test subjects tend to risk contributing with half the amount they have received to the sum, which on being doubled, turns to 100%; which is then split in four, resulting in 25% for each and, in the end, you only lose one quarter. If everyone does the same, you actually double what you put in. It is also settled beforehand how many times this game will happen — about 10 or 20. But, towards the end, participants have the tendency to contribute with less money, expecting to gain from what others put in. So, in the beginning, willing to earn the trust of others, they contribute more, but then they progressively contribute less and less, to benefit more before the game ends.

We can make the game even more complicated. I may want to punish someone for not contributing at all or, on average, contributing with less than they should. That, however, will take a personal cost. If I want to punish someone by 70%, I have to chip in with the remaining 30% to sum up the 100% of a total amount which is returned to the scientist. In order to punish I have to pay, and to do that I have to be effectively willing to fulfill my intent. What happens is that people start to punish those who contribute with less, in such a way that it leads the culprit to increase his contributions. The so changed game — it has been verified experimentally — makes the contributions rise up to 100%, both for the duplication of the sum as well as for avoiding punishment. This happens even if there are several groups, and each person keeps changing groups and never re-encounters old game partners — the punishment behavior remains the same. What we conclude is that it is not an educational punishment, because groups are never reconstructed, but that there is in us a tendency to punish the infractor, to make him comply. One can also argue that something makes altruism rules stick in a large group. We do not want to educate a particular individual, but only to keep a general group culture. However, it is also by vengeance, and there exists a retaliation mechanism which is strong up to the point where, even when one changes group, that «trait» still functions. It is in these terms that the origin of the vengeance emotion is explained.

There is yet another game we should mention, said the «ultimatum» game. Someone comes up with 100 dollars and gives them to me. I offer 70 of those 100 dollars to another individual. If he accepts my offer, knowing I get the remaining 30 dollars, that is how the split takes place. If he does not accept we both get nothing. I could have the tendency to offer 1 dollar to the other, which would accept it thinking: «if it weren't for this game I would gain nothing, so I'll take this small offer, it's a gain for me anyway. I'll take it». But the majority of people do not accept such a lowly offer, and accepts only if the offer is between 40 and 50 dollars. Most of the players offering money usually also offer about that amount, keeping a little more than what they offer. If one offers less than 40 others tend to reject the offer, the justification going along these lines: «under such circumstances I don't want anything». They are «irrationally» willing to let go of a pure additional gain because, rationally, there should be no concern about what the offeror gains or not. Due to our innate tendency, we already know that social games are repeated. Even if you tell people that the game is going to be played only once, our genetic — and memetic — programming refuses to accept small amounts, just to make others not be so selfish, to keep the so called altruistic

group culture. However, when the percentage is decided by a computer — someone gives me 100 dollars and I just press a button to hear the money distribution instructions given by the computer — the other person usually accepts anything because he considers there was an impersonal decision maker who will not be influenced.

In the Machiguenga tribes of Amazon, and Papua New Guinea, people offer more. In western civilization the common practice is to accept 45%, in the Amazon people stick with only 26% and in Papua New Guinea they offer above 50%, because they are already used to give away the pigs as we have discussed previously, and this is where the concepts of social respect and social debt come into play. The game is being considered in a broader context, not just as an isolated experiment, as in the commercial transactions we are used to.

It is curious that, in these experimental games, certain phenomena are concocted to bring up aspects that change with age. One example is self-esteem, responsible for the rejection of a low offer in such a way that nobody gets tagged as the one who settles for little and to whom we need not offer that much. Self-esteem drives us, even in the absence of consciousness of a clear rejection strategy, like when the waiter at the restaurant rejects a very small tip.

Let us return to the sum game with the punishment possibility. Even when circulating amongst several groups, never re-encountering old game partners, and excluding the notion of education, punishment still persists. From the game theory point of view, vengeance is the feeling that leads to punishment and it is extremely useful, thereby memetically surviving.

Many of our emotions are deceit strategies because, from the moment they exist, and aware of their characteristic of being non dissociable from the human being, we can try to trigger them and move on to the next level: the emotional game. But, first, the game has to be seen from the point of view of survival by the best use of strategy. Those who get resources survive, those who do not go extinct. However, the essence of the game is that both can win or both can lose. The computational simulations, in greatly sophisticated games with a high topological complexity, show what is the best strategy that survives, multiplies and is stable, and they show that altruism and cooperation emerge and spread, under very broad conditions.

We cannot ignore mutations. At any given time a strategy can be altered and the individual may try others. There can be an evolutionary mutation in which people perform differently in their participation tasks. When faced with mutations, we no longer have before us classic games in which there can be infinitely many immutable equilibrium points, those in which each equilibrium point persists because any small change in the payoff of the game move with the ongoing strategy does not bring about new benefits nor additionally avoids harm. By definition, all these are equilibrium points. But it will not be a classic equilibrium if there will occur mutations that give rise to unexpected strategy variants. In such cases, if the equilibrium point still persists anyway, the strategy is said to be evolutionary stable, thus generalizing the concept of Nash Equilibrium. Evolutionary Game Theory studies such circumstances.

Even in simulated situations, where there is neither the education nor the learning factors, and in which individuals know that, in each game, they will not match the same old partners, they still execute certain strategies which are already imprinted in our brains and in the way we behave. It has nothing to do with the consciousness of wanting to influence, but with something already related to feelings and deeper cognitions.

Actually, the structure of games, the structure of repetitions and of encounters — the choosing or not of whom we match with, whose place we go to and whom we welcome in our house — all that game, all those games space possibilities, have been played during the learning of our species. There is a genetic or memetic learning that develops those frames of reference. Evolution itself is a strategy game.

Rationalization itself might be part of a strategy. An individual, if he wants to dispute something, rationalizes. Guilt tends to rise when he begins to feel that he might be caught and, as I have said before, that guilt might ultimately lead him to a pre-empting confession. Laughter itself is a strategy that evolved from a display of aggressiveness into a strategy to appease the other. We can look into the physiological characteristics of laughter and try to understand which were the first motivations that made our body to physically adapt to the production of certain substances and subsequent consequences.

Games can be treated as typifications of social organization, as if they were logical equations subject to theoretical and/or experimental evaluation. Knowledge of them is itself part of a game in which we all can win, because its derived conclusions supply a better knowledge of the reality from which we could move on towards a better game. There are cumulative sets of results that make us move on to the next level, our knowledge grows and we can benefit from that ensemble. This type of game has to be played with a very rigorous tactic.

We began by saying there is a combinatorial game composed by genes, and that certain stable structures can get more complex and give rise to a generating combination, to a nervous system with the capability of reproducing ideas, as well as of modeling external features, of making retrospectives, of aiming for predictions of the future, etc. How did cognition reach this point through evolution? Today science already has the ability to prognosticate and to give some thought-through answers to these questions. But we can see that, at a large scale, the name of the game is to survive and replicate, perhaps only by mere copy — because only by achieving that genetic/memetic reproduction there is a future where the game continues to make sense. Only this way the game can become even more complicated, and it is only by increasing complication that it is possible to aspire to be a better player and to take advantage of the worst players, making them evolve into better ones, in such a way as to attain maximum common benefits, unachievable without generalized cooperation.

We are immersed in games, strategies, coexistence of evolutionary strategies. While psychology and psychoanalysis focus very much on the individual's past, they never truly looked into the specie's past. Games create hidden ghosts in terms of certain cognitions, since in the evolutionary game there can be a potential advantage in deceiving the other, even in deceiving oneself. In terms of

evolutionary competition there coexist, however, other variants like the games where everyone can win, but involving intra and intergroup altruism problems, always subject to individuals' emotions. The behaviors of human societies, in terms of phenomena, emergent or otherwise, are so deeply underlying that often not even the actors themselves are aware of them, and behaviors express themselves because that corresponds to a cognitive evolution coded in the collective unconscious, which is circumstantially actionable. The very emotions, which are usually seen as opposed to rationality, end up possibly being compiled strategies, that survived as such in a certain type of game. We can then well imagine how emotional and sexual behaviors in general accomplish hidden purposes.

We can easily understand, through computer simulations of competing populations and their respective planning, that the winning strategies change over time. They change in accordance with the probabilities of encountering an opponent with a different tactic, or when we have encountered them before we reorganize ourselves to deal with individuals who share our «tricks» in a manner as to be able to cooperate in order to achieve better results. Here we discover social organization, and it is by these means that exterior signals come about — among them, language reveals itself as one of the most important — which are identifiers of social types, including the facet of the opportunist who pretends to cooperate to take some advantage. Strategies, however, have also progressed to detect so-called opportunists, there being those who maintain that the brain evolved as the result of a complex adaptation to the social system, with its vicissitudes and abusers, since in order to live in this natural world we do not need a very sophisticated brain, as many animals so prove.

It is quite interesting that nowadays all these subjects have begun to be studied via mathematical models and algorithms, using the computer to simulate strategies and allowing us to understand the emergent phenomena. In the current state, we begin to instill in the players of these games more flexible cognitive abilities, coming from the Artificial Intelligence domain, like the above mentioned Intention Recognizer, thus allowing the achievement of new levels of sophistication of game models and of the study of cooperative success, encompassing individual and group tolerance.

## **5 Group Altruism/Opportunism**

As aforementioned, whenever there is altruism there is opportunism. Let us imagine, for example, what can happen with the European social model. It has guaranteed pensions, it has a health system for everyone —society has created altruistic mechanisms. Obviously, there are people filing fake diseases, taking advantage of each and every loophole to unduly benefit from that very altruism. We must not only repress the transgressors, but also those who are in charge of repressing them and do not do so: the corrupts. There are individuals who are bribed by the ill-intended because, obviously, the transgressor uses part of his benefit to pay the corrupt fiscal who “closes his eyes”. This game between altruism and parasitism is inevitable — as inevitable as the power of gravity.

Because there are always mutations that promote altruism, there are mutations that create opportunists who will take advantage of the former. And often they do thrive.

It has been proved, in mathematical models of evolutionary games, that group selection has to do with the memetic pressure for compliance with the rules of the game. But such compliance will only take place if those who do not labor to enforce it, whenever they have the opportunity to do so, are penalized as well. Such is the case since caring and penalizing have a cost that, if we can we will avoid when compliance does not affect us directly, unless we ourselves are penalized for doing nothing.

We have this personal experience in traffic — many times we repress or preventively impede others from doing some maneuver, when faced with a car which we do not know whether it will be trying or not to get forcefully into the queue, thereby violating a traffic law. Playing it safe, we impede this wayward possibility — we are then enforcing the rules of the system, saying to opportunists, manifest or potential, that there are rules and violators are not tolerated.

In terms of human societies, the above means that altruism cannot be for everyone. A person is altruistic towards the group he/she belongs to. It is evident that, when you belong to a group there can be outside groups that will try to take advantage of your own.

Imagine a human group of the Upper Paleolithic or Lower Neolithic that farms, domesticates cattle, has corn already under control, even achieving some genetic improvements — using the mechanisms of natural selection to, in an artificial manner, improve the species of its cattle or grain. This group lives with a certain wealth, it has its society set up, it has a reproductive cycle, their women are able to generate healthy children. It is obvious that another group will always be tempted to attack this group to take the resources it built up. Because resources correspond to an investment, the social organization too corresponds to an investment, and those who have not made that investment will do better if they steal.

There are mechanisms for groups to get along, even on account of their having to exchange genes among themselves. Thus, genes from one group are exported to others and interchange takes place. There are mechanisms of commercial trade that are very important. Commercial trade builds trust relations, and this brings about the problem of credibility: After all, who is the partner I am playing with? After all, which individuals do I prefer to make transactions with? But always there are pirates, and there are still pirates in the seas even today, where there are no law enforcers.

Reciprocal altruism occurs towards a group we know is behaving according to certain rules. We are not going to be altruistic towards strangers, especially if they are with another group whose rules of the game we completely ignore. We ignore if they are deceiving us, if they have second intentions, and what are their behaviors as a group towards others. It is required to build a previous trust, which must undergo the declaration of identities and intentions, and the holding of coherence.

We have previously mentioned individual natural selection, more related to genes. But, when individuals group into units, we can also consider that there is group selection. However, claiming it is not enough, it is necessary to prove it mathematically and with computer simulations. For a long time this notion of group selection has been rejected, except for very specific and rare circumstances. In truth, it can be proven by mathematical simulation that, in fact, individuals of groups had to exchange genes in order to diversify their «pool» for to avoid hereditary diseases. Because, having a double helix, the gene in front of another with a malefic mutation can correct it. However, if both are equal, a hereditary disease manifests itself. That is why there are so many incest prohibiting rules, such rules having appearing spontaneously. Groups who did live by them did not resist the illnesses resulting from «inbreeding». It is necessary to exchange genes. The group must keep its borders open at all times.

Only very recently have we began to look at more sophisticated models and to make memes part of simulations. Indeed, memes code for algorithms, social routines, which afford a behavioral identity and unity to the group, and are interpretable in varied and overlapping relational allegiances with their distinct mechanisms.

The other side of the coin of group unity is, naturally, competition amongst groups. For example, the group can, via its memetic-religious ideology — through a shared divine bonding —, be more fierce, more aggressive, and take all the opportunities to slaughter others. Group behavior, inward and outward, is now determined as well by the memes, a reproductive mechanism which is faster and more flexible than the genetic one — as we said before, genes take a generation to propagate while memes take only the time of a culture sharing act. What survives is the memetic combination of the group, confronted with the other groups and attending strategies. Of course genes are still there. And there stands the global problem of how these two reproductive levels co-inhabit.

We now begin entering *terra incognita*, the issue of the interaction between memes and genes. Because, on the one hand, on some occasions, they are antagonists, but on other occasions they have an interest in cooperating. Memes are relatively recent in evolutionary terms. It was our species who took them to a progressive refinement, only possible because we have language. Language — and it does not necessarily have to be the spoken word, it can be a sign language — is the form *par excellence* of transmitting memes. This memetic reproduction, in the societies we live in today, tends to say that individuals should be treated equally no matter what their genetic combination is (no matter the color of the skin, with or without disabilities, etc.). Our own memetic culture states that genetic difference does not matter — everyone is memetically treated *a priori* in the same way. This is the case in some societies, in others it can be different. We can see that memes themselves already can control genes: by genetic manipulation they can handle them, in a good sense, curing hereditarily transmissible diseases; or also, in a bad sense, given that they could be empowered in eugenics and race improvement programs.

We are still in the beginning of knowing how they work in articulation, these two reproductive mechanisms. From a computational perspective, bottom line,

they can be seen in terms of co-present strategies — and it does not matter whether their underlying support is biochemical, or if it is the C++ programming language, or any other. In abstract, what we are studying are certain functionalities in co-presence. But one can say that human evolution is getting ever more memetic (Richerson & Boyd, 2006).

## 6 Complex Networks of Mindful Entities: A New Research Domain

With our research on these networks we intend to understand, and explain, how some social collective behaviors emerge from the cognitive capabilities of individual agents, in communities where said individuals are nodes of complex adaptive networks, which self-organize as a result of the referred cognition of individual agents. Consequently, we need to investigate which cognitive abilities have an impact on the emergence of properties of the population and, as a result, which cognitive abilities determine the emergence of which specified social collective behaviors. Hence, the key innovation consists in the articulation of two distinct levels of simulation, individual and social, and in their combined dynamics. This needs to be reified both at the modeling level as well as at the computational implementation one.

Biological evolution is characterized by a set of highly braided processes, which produce a kind of extraordinarily complex combinatorial innovation. A generic term frequently used to describe this vast category of spontaneous, and weakly predictable, order generating processes, is «emergence». This term became a kind of signal to refer the paradigms of research sensitive to systemic factors. Complex dynamic systems can spontaneously assume patterns of ordered behaviors which are not previously imaginable from the properties of their composing elements nor from their interaction patterns. There is unpredictability in self-organizing phenomena — preferably called «evolutionary» —, with considerably diverse and variable levels of complexity.

What does emerge? The answer is not something defined but instead something like a shape, or pattern, or function. The concept of emergence is applicable to phenomena in which the relational properties predominate over the properties of composing elements in the determination of the ensemble's characteristics. Emergence processes appear due to configurations and topologies, not to properties of elements (Deacon, 2003).

As we have remarked before, two hundred years after the birth of Darwin, and 150 after the *On the Origin of Species*, several fundamental questions about evolution still remain unanswered. The problem of evolution of cooperation and of the emergence of collective action — cutting across areas as diverse as Biology, Economy, Artificial Intelligence, Political Science, or Psychology — is one of the greatest interdisciplinary challenges science faces today. To understand the evolutionary mechanisms that promote and keep cooperative behavior is all the more complex as increasingly intricate is the intrinsic complexity of the partaking individuals. «Complexity» refers to the study of the emergence of collective

properties in systems with many interdependent components. These components can be atoms or macro molecules in a physical or biological context, and people, machines or organizations in a socioeconomic context.

This complexity has been explored in recent works, where it is shown, amongst several other properties, that the diversity associated with structures of interaction, of learning and reproduction of a population, is determinant for the choices of agents and, in particular, to the establishment of cooperation actions (Santos *et al.*, 2006, 2008). These studies were based on the frame of reference provided by Evolutionary Game Theory (Maynard-Smith, 1982) — alluded to before — and by the theories of Science of Networks (Dorogotsev & Mendes, 2003), combining instruments for modeling multi-agent systems and complex adaptive systems.

«Egotism» concerns the logic behind the unending give-and-take that pervades our societal lives. It does not mean blind greed, but instead an informed individual interest. Thus, «The evolution of cooperation» has been considered one of the most challenging problems of the century. Throughout the ages thinkers have become fascinated by the issue of self-consideration versus “the other”-consideration, but the use of formal models and experimental games is relatively recent. Since Robert Trivers (Trivers, 1971) introduced the evolutionary approach to reciprocity, games have served as models to explore the issue.

The modeling of artificial societies based on the individual has significantly expanded the scope of game theory. Societies are composed by fictitious subjects, each equipped with a strategy specified by a program. Individuals meet in randomized pairs, in a joint iterated game.

The comparison of accumulated rewards is used to update the population: the most successful individuals produce more offspring, which inherit their strategy. Alternatively, instead of inheriting strategies, new individuals may adapt by copying, from known individuals, the strategies that had best results. In both cases, the frequency of each strategy in the population changes over time, and the ensemble may evolve towards a stable situation. There is also the possibility of introducing small mutations in minority, and study how they spread.

Evolutionary Game Theory (EGT) is necessary to understand the why and the how of what it takes for agents with individual interests to cooperate for a common weal. EGT emphasizes the deterministic dynamics and the stochastic processes. Repeated interactions allow the exploration of direct reciprocity between two players (Sigmund, 2010).

In the EGT approach the most successful strategies become more frequent in the population. Kinship, neighborhood relationships, and individual differences, may or may not be considered. In indirect reciprocity (Nowak & Sigmund, 2005), players interact at most once, but they have knowledge of their partners' past behavior. This introduces the concern with reputation, and eventually with moral judgment (Pacheco & Santos & Chalub, 2006; Pereira & Saptawijaya, 2011; Han & Saptawijaya & Pereira, 2012).

The strategies based on the evaluation of interactions between third parties allow the emergence of kinds of cooperation that are immune to exploitation, because then interactions are channeled just to those who cooperate. Questions of justice and trust, with their negative (punishment) and positive (help) incentives,



are fundamental in games with large diversified groups of individuals gifted with intention recognition capabilities. In allowing them to choose amongst distinct behaviors based on suggestive information about the intentions of their interaction partners, they are, in turn, influenced by the behavior of the individual himself, and are also influenced by the tolerance to error and to noise in the communication. One hopes understanding these capabilities can be transformed into mechanisms for spontaneous organization and control of swarms of autonomous robotic agents.

With this objective, we have studied the way players' strategies adapt in populations involved in cooperation games. We used the techniques of EGT and considered games such as the «Prisoner's Dilemma» and «Stag Hunt» successively repeated, and showed how the actors participating in repeated iterations with these games can benefit from having the ability to recognize the intentions of other actors, leading to an evolutionary stable increase in cooperation (Anh & Pereira & Santos, 2011, 2011a, 2012, 2012a), compared to extant best strategies.

Intention recognition is implemented using «Bayesian Networks» (BN) and taking into account the information of current signals of intent, as well as the trust and tolerance built from previous plays. We experimented with populations with different proportions of diverse strategies in order to calculate, in particular, what is the minimum fraction of individuals with Intention Recognition for cooperation to emerge, invade, prevail, and persist.

## 7 Directions for the Future

The fact that, in a networked population, individuals can have more cognitive capabilities and dynamically choose their behavior rules — instead of acting from a predetermined set — gives the system a much richer and realistic dynamics, worth exploring. Within the scope of this new paradigm, individuals must be able to hypothesize, to look at possible futures, to probabilistically prefer, to deliberate, to send and respond to signals, to take into account history and trust, to form coalitions, to adopt and fine tune game strategies.

Actually, the study of those properties that emerge from populations in complex networks still needs to further investigate the cognitive core of each of the social atoms. Given the plethora of possibilities in the modeling of cognitive capabilities, we must identify the intrinsic characteristics which, solely by themselves, provide the most prominent individual behavior, and are conducive to an emergent collective behavior which cannot be anticipated, but is cooperative and tolerant. It is required to consider limiting the number of available parameters, in such a way as to render the study treatable, and also to make it implementable in future «robots», in the engineering domain and not just in the simulation domain.

All things considered, one should take into account different types of individual and social cooperation dynamics, whether deterministic or stochastic, and use N-people interactions modeled in terms of evolutionary games that constitute metaphors of the social dilemmas of cooperation. It seems to us that Intention Recognition, and its use in the scope of tolerance, is a foundational cornerstone

where we should begin at, naturally followed by the capacity to establish and honor «commitments», as a tool towards the successive construction of collective intentions and social organization (Searle, 2010).

## 8 Coda

Evolutionary Psychology and Evolutionary Game Theory provide a theoretical and experimental framework for the study of social exchanges, where tolerance towards the inside of a group and discrimination and intolerance towards the outside of the group are the two sides of the same coin. The strategic recipe «love thy neighbor» often paradoxically contains the genesis of hatred and war, because neighbor refers to the «tribe», and the gods are referees on our side.

Recognition of someone's intentions, which may include imagining the recognition others have of our own intentions, and comprise some error tolerance, can lead to evolutionary stable win/win equilibriums within groups of individuals and amongst groups. The manifestation of intentions is a facilitator in that respect. Additionally, by means of joint objectives under commitment, one can promote the inclusion of heretofore separate groups into a more global one. The overcoming of intolerance shall benefit from both these levels of manifest interaction.

We have argued that the study of these issues in minds with machines has come of age and is ripe with research opportunities, and have also communicated some of the published inroads we have achieved with respect to intention recognition and tolerance in the evolutionary game theory context.

**Acknowledgements.** This paper is the author's updated version, in English, of its original in Portuguese, titled "Tolerância Evolucionária", published in "Revista Portuguesa de Psicanálise", 30(2) 117-147, 2010. I thank their permission to publish its translation here. Moreover, I thank Francisco C. Santos for his expert help in improving the original.

## References

- Axelrod, R.: *The Evolution of Cooperation*. Basic Books, Cambridge (1984)
- Bowlby, J.: *Attachment and Loss: Attachment*, vol. 1. Penguin Books, London (1971)
- Buss, D.M.: *The Handbook of Evolutionary Psychology*. John Wiley & Sons Inc., New York (2005)
- Damasio, A.: *Self Comes to Mind*. William Heinemann, Portsmouth (2010)
- Deacon, T.W.: *The Hierarchic Logic of Emergence: Untangling the Interdependence of Evolution and Self-Organization*. In: Weber, H.W., Depew, D.J. (eds.) *Evolution and Learning: The Baldwin Effect Reconsidered*. MIT Press, Cambridge (2003)
- Dennett, D.C.: *Darwin's Dangerous Idea – Evolution and the Meanings of Life*. Simon & Schuster, New York (1995)
- Dorogotsev, S.N., Mendes, J.F.F.: *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford (2003)
- Dunbar, R.I.M., Barrett, L.: *Oxford Handbook of Evolutionary Psychology*. Oxford University Press, Oxford (2007)

- Gangestad, S.W., Simpson, J.A.: *The Evolution of Mind*. The Guilford Press, New York (2007)
- Han, T.A., Pereira, L.M.: State-of-the-Art of Intention Recognition and its use in Decision Making. Submitted to *AI Communications* (2012)
- Han, T.A., Pereira, L.M.: Context-dependent Incremental Intention Recognition via Bayesian Network Model Construction. Draft (2012a)
- Han, T.A., Pereira, L.M., Santos, F.C.: Intention Recognition Promotes The Emergence of Cooperation. *Adaptive Behavior* 19(3), 264–279 (2011)
- Han, T.A., Pereira, L.M., Santos, F.C.: The role of intention recognition in the evolution of cooperative behavior. In: *Proc Intl. Joint Conf. on Artificial Intelligence (IJCAI 2011)*, Barcelona, Spain, pp. 1684–1689 (July 2011a)
- Han, T. A., Pereira, L.M., Santos, F.C.: The emergence of commitments and cooperation. Accepted in *AAMAS 2012. ACM Proceedings*, Valencia, Spain (June 2012)
- Han, T.A., Pereira, L.M., Santos, F.C.: Corpus-based Intention Recognition Leads to the Emergence of Cooperative Behavior. Submitted to *Artificial Life* (2012a)
- Han, T.A., Saptawijaya, A., Moniz Pereira, L.: Moral Reasoning under Uncertainty. In: Bjørner, N., Voronkov, A. (eds.) *LPAR-18 2012. LNCS(LNAD)*, vol. 7180, pp. 212–227. Springer, Heidelberg (2012)
- Hölldobler, B., Wilson, E.O.: *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies*. W.W. Norton, New York (2009)
- Kirkpatrick, L.A.: *Attachment, Evolution, and the Psychology of Religion*. The Guilford Press, New York (2005)
- Laland, K.N., Brown, G.R.: *Sense & Nonsense: Evolutionary Perspectives on Human Behaviour*. Oxford University Press, Oxford (2002)
- Maynard-Smith, J.: *Evolution and the Theory of Games*. Cambridge University Press, Cambridge (1982)
- Mithen, S.: *The Prehistory of Mind*. Thames & Hudson Ltd, London (1996)
- Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity. *Nature* 437, 1291–1298 (2005)
- Pacheco, J.M., Santos, F.C., Chalub, F.A.: Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput. Biol.* 2, e178 (2006)
- Pereira, L.M.: Evolving Towards Evolutionary Epistemology. *International Journal of Reasoning-based Intelligent Systems* 1(1-2), 68–76 (2009)
- Pereira, L.M., Saptawijaya, A.: Modelling Morality with Prospective Logic. In: Anderson, M., Anderson, S.L. (eds.) *Machine Ethics*, pp. 398–421. Cambridge University Press (2011)
- Pereira, L.M., Han, T.A.: Intention Recognition with Evolution Prospection and Causal Bayesian Networks. In: Madureira, A., Ferreira, J., Vale, Z. (eds.) *Computational Intelligence for Engineering Systems: Emergent Applications*, pp. 1–33. Springer, Berlin (2011)
- Platak, S.M., Keenan, J.P., Shackelford, T.K.: *Evolutionary Cognitive Neuroscience*. The MIT Press, Cambridge (2007)
- Richerson, P.J., Boyd, R.: *Not By Genes Alone: How Culture Transforms Human Evolution*. The University of Chicago Press, Chicago (2006)
- Rizzolatti, G., Sinigaglia, C.: *Mirrors in the Brain: How our Minds Share Actions and Emotions*. Oxford University Press, Oxford (2007)
- Santos, F.C., Pacheco, J.M., Lenaerts, T.: Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl. Acad. Sci. U S A* 103, 3490–3494 (2006)

- Santos, F.C., Santos, M.D., Pacheco, J.M.: Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454, 213–216 (2008)
- Searle, J.: *Making the Social World: The Structure of Human Civilization*. Oxford University Press, Oxford (2010)
- Shennan, S.: *Genes, Memes and Human History – Darwinian Archaeology and Cultural Evolution*. Thames & Hudson Ltd, London (2002)
- Sigmund, K.: *The Calculus of Selfishness*. Princeton University Press, Princeton (2010)
- Skyrms, B.: *Evolution of the Social Contract*. Cambridge University Press, Cambridge (1996)
- Skyrms, B.: *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge (2004)
- Skyrms, B.: *Signals – Evolution, Learning, & Information*. Oxford University Press, Oxford (2010)
- Trivers, R.L.: The evolution of reciprocal altruism. *Quarterly Review of Biology* 46, 35–57 (1971)
- von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press, Princeton (1944)
- Wright, R.: *NonZero – The Logic of Human Destiny*. Vintage, New York (2001)