STUDIES
IN APPLIED
PHILOSOPHY,
EPISTEMOLOGY
AND
RATIONAL
ETHICS

Lorenzo Magnani (Ed.)

SAPERE

Model-Based Reasoning in Science and Technology

Theoretical and Cognitive Issues

Springer

Studies in Applied Philosophy, Epistemology and Rational Ethics

Volume 8

For further volumes:
http://www.springer.com/series/10087

About This Series

Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE) publishes new developments and advances in all the fields of philosophy, epistemology, and ethics, bringing them together with a cluster of scientific disciplines and technological outcomes: from computer science to life sciences, from economics, law, and education to engineering, logic, and mathematics, from medicine to physics, human sciences, and politics. It aims at covering all the challenging philosophical and ethical themes of contemporary society, making them appropriately applicable to contemporary theoretical, methodological, and practical problems, impasses, controversies, and conflicts. The series includes monographs, lecture notes, selected contributions from specialized conferences and workshops as well as selected PhD theses.

Lorenzo Magnani
Editor

Model-Based Reasoning in Science and Technology

Theoretical and Cognitive Issues

Springer

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

ISSN 2192-6255 ISSN 2192-6263 (electronic)
ISBN 978-3-642-37427-2 ISBN 978-3-642-37428-9 (eBook)
DOI 10.1007/978-3-642-37428-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013941542

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume is a collection of selected papers that were presented at the international conference *Model-Based Reasoning in Science and Technology. Theoretical and Cognitive Issues* (MBR012_Italy), held at the Fondazione Mediaterraneo, Sestri Levante, Italy, in June 2012.

A previous volume, *Model-Based Reasoning in Scientific Discovery*, edited by L. Magnani, N. J. Nersessian, and P. Thagard (Kluwer Academic/Plenum Publishers, New York, 1999; Chinese edition, China Science and Technology Press, Beijing, 2000), was based on the papers presented at the first Model-Based Reasoning international conference, held at the University of Pavia, Pavia, Italy, in December 1998. Other two volumes were based on the papers presented at the second Model-Based Reasoning international conference, held at the same place in May 2001: *Model-Based Reasoning. Scientific Discovery, Technological Innovation, Values*, edited by L. Magnani and N. J. Nersessian (Kluwer Academic/Plenum Publishers, New York, 2002) and *Logical and Computational Aspects of Model-Based Reasoning*, edited by L. Magnani, N. J. Nersessian, and C. Pizzi (Kluwer Academic, Dordrecht, 2002). Another volume, *Model-Based Reasoning in Science and Engineering*, edited by L. Magnani (College Publications, London, 2006), was based on the papers presented at the third Model-Based Reasoning international conference, held at the same place in December 2004. The volume *Model-Based Reasoning in Science and Medicine*, edited by L. Magnani and L. Ping (Springer, Heidelberg/Berlin 2006), was based on the papers presented at the fourth Model-Based Reasoning conference, held at Sun Yat-sen University, Guangzhou, P. R. China. Finally, the volume *Model-Based Reasoning in Science and Technology, Abduction, Logic, and Computational Discovery*, edited by L. Magnani, W. Carnielli, and C. Pizzi (Springer, Heidelberg/Berlin 2010), was based on the papers presented at the fifth Model-Based Reasoning conference, held at the University of Campinas, Campinas, Brazil, in December 2009.

The presentations given at the Sestri Levante conference explored how scientific thinking uses models and explanatory reasoning to produce creative changes in theories and concepts. Some speakers addressed the problem of model-based reasoning in technology, and stressed issues such as the relationship between science and technological innovation. The study of diagnostic, visual, spatial, analogical, and temporal reasoning has demonstrated that there are many ways of performing intelligent and creative reasoning that cannot be described with the help of traditional notions of reasoning such as classical logic. Understanding the contribution of modeling practices to discovery and conceptual change in science and in other disciplines requires expanding the concept of reasoning to include complex forms of creativity that are not always successful and can lead to incorrect solutions. The study of these heuristic ways of reasoning is situated at the crossroads of philosophy, artificial intelligence, cognitive psychology, and logic, that is, at the heart of cognitive science. There are several key ingredients common to the various forms of model-based reasoning. The term “model” comprises both internal and external representations. The models are intended as interpretations of target physical systems, processes, phenomena, or situations. The models are retrieved or constructed on the basis of potentially satisfying salient constraints of the target domain. Moreover, in the modeling process, various forms of abstraction are used. Evaluation and adaptation take place in light of structural, causal, and/or functional constraints. Model simulation can be used to produce new states and enable evaluation of behaviors and other factors. The various contributions of the book are written by interdisciplinary researchers who are active in the area of modeling reasoning and creative reasoning in logic, cognitive science, and science and technology; the most recent results and achievements about the topics above are illustrated in detail in the papers.

The editor expresses his appreciation to the members of the Scientific Committee for their suggestions and assistance:—Atocha Aliseda, Instituto de Investigaciones Filosoficas, Universidad Nacional Autónoma de Mexico (UNAM), Mexico—Emanuele Bardone, Institute of Informatics, University of Tallinn, Estonia—Silvana Borutti, Department of Humanities, Philosophy Section, University of Pavia, Italy—Otàvio Bueno, Department of Philosophy, University of Miami, Coral Gables, USA—Mirella Capozzi, Department of Philosophy, University of Rome La Sapienza, Rome, Italy—Walter Carnielli, Department of Philosophy, Institute of Philosophy and Human Sciences, State University of Campinas, Brazil—Claudia Casadio, Department of Psychology, University of Chieti-Pescara, Italy—Carlo Cellucci, Department of Philosophy, University of Rome La Sapienza, Rome, Italy—Sanjay Chandrasekharan, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA—Roberto Cordeschi, Department of Philosophy, University of Rome La Sapienza, Rome, Italy—Roberto Feltrero, Department of Logic, History and Philosophy of Science at UNED (Spanish Open University), Madrid, Spain—Steven French, Department of Philosophy, University of Leeds, Leeds, UK—Marcello Frixione, Department of Communication Sciences, University of Salerno, Italy—Dov Gabbay, Department of Computer Science, King’s College, London, UK—Marcello Guarini, Department of Philosophy, University of Windsor, Canada—Ricardo Gudwin, Department of Computer Engineering and Industrial Automation, the School of Electrical Engineering and Computer Science, State University of Campinas, Brazil—Viorel Guliciuc, Stefan cel Mare University, Suceava, Romania—Albrecht Heeffer, Center for History of Science, Ghent University, Belgium—Michael Hoffmann, School of Public Policy, Georgia Institute of Technology, Atlanta, USA—Decio Krause, Departamento de Filosofia, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil—Michael Leyton, Psychology Department, and DIMACS, Center for Discrete Mathematics, and Theoretical Computer Science, Rutgers University, USA—Ping Li, Department of Philosophy, Sun Yat-sen University, Guangzhou, P. R. China—Giuseppe Longo, CREA, CNRS and École Polytechnique, Paris, France—Angelo Loula, Department of Exact Sciences, State University of Feira de Santana, Brazil—Shangmin Luan, Institute of Software, The Chinese Academy of Sciences, Beijing, P. R. China—Rossella Lupacchini, University of Bologna, Bologna, Italy—Joke Meheus, Vakgroep Wijsbegeerte, Universiteit Gent, Gent, Belgium—Woosuk Park, Humanities and Social Sciences, KAIST, Guseong-dong, Yuseong-gu Daejeon, South Korea—Claudio Pizzi, Department of Philosophy and Social Sciences, University of Siena, Siena, Italy—Demetris Portides, Department of Classics and Philosophy, University of Cyprus, Nicosia, Cyprus—Joao Queiroz, Institute of Arts and Design. Federal University of Juiz de Fora, Brazil—Shahid Rahman, U.F.R. de Philosophie, Université Lille 3, Villeneuve d'Ascq, France—Oliver Ray, Department of Computer Science, University of Bristol, Bristol, UK—Colin Schmidt, Institut d'Informatique Claude Chappe, University of Le Mans, France—Gerhard Schurz, Institute for Philosophy, Heinrich-Heine University, Germany—Cameron Shelley, Department of Philosophy, University of Waterloo, Waterloo, Canada—Chris Sinha, Centre for Cognitive Semiotics, Lund University, Lund, Sweden—Nik Swoboda, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain—Paul Thagard, Department of Philosophy, University of Waterloo, Waterloo, Canada—Barbara Tversky, Department of Psychology, Stanford University, Stanford, USA and Teachers College, Columbia University, New York, USA—Ryan D. Tweney, Emeritus Professor of Psychology, Bowling Green State University, Bowling Green, USA—Riccardo Viale, Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy and Italian Cultural Institute of New York (Director), USA—John Woods, Department of Philosophy, University of British Columbia, Canada, and also to the members of the local scientific committee: Emanuele Bardone (University of Tallinn), Tommaso Bertolotti (University of Pavia), and Elena Gandini (Across Events, Pavia).

Special thanks to Tommaso Bertolotti for his contribution in the preparation of this volume. The conference MBR012_Italy, and thus indirectly this book, was made possible through the generous financial support of the MIUR (Italian Ministry of the University) and of the University of Pavia. Their support is gratefully acknowledged. The preparation of the volume would not have been possible without the contribution of resources and facilities of the Computational Philosophy Laboratory and of the Department of Humanities, Philosophy Section, University of Pavia.

Several papers concerning model-based reasoning deriving from the previous conferences MBR98 and MBR01 can be found in Special Issues of Journals: in *Philosophica*: Abduction and Scientific Discovery, 61(1), 1998, and Analogy and

Mental Modeling in Scientific Discovery, 61(2) 1998; in *Foundations of Science*: Model-Based Reasoning in Science: Learning and Discovery, 5(2) 2000, all edited by L. Magnani, N.J. Nersessian, and P. Thagard; in *Foundations of Science*: Abductive Reasoning in Science, 9, 2004, and Model-Based Reasoning: Visual, Analogical, Simulative, 10, 2005; in *Mind and Society*: Scientific Discovery: Model-Based Reasoning, 5(3), 2002, and Commonsense and Scientific Reasoning, 4(2), 2001, all edited by L. Magnani and N.J. Nersessian. Finally, other related philosophical, epistemological, and cognitive oriented papers deriving from the presentations given at the conference MBR04 have been published in a Special Issue of the *Logic Journal of the IGPL*: Abduction, Practical Reasoning, and Creative Inferences in Science, 14(1), (2006) and have been published in two Special Issues of *Foundations of Science*: Tracking Irrational Sets: Science and Technology, Ethics, and Model-Based Reasoning in Science and Engineering, 13(1) and 13(2) (2008), all edited by L. Magnani. Other technical logical papers presented at MBR09 BRAZIL have been published in a Special Issue of the *Logic Journal of the IGPL*: Formal Representations in Model-Based Reasoning and Abduction, 29(2), (2012), edited by L. Magnani, W. Carnielli, and C. Pizzi.

Other more technical formal papers presented at (MBR012 ITALY) will be published in a special issue of the *Logic Journal of the IGPL*, edited by L. Magnani.

Finally, the present book also includes a paper *How to Learn Abduction from Animals? From Avicenna to Magnani*, that Woosuk Park has devoted to the 60th birthday of the chair of the conference.

Pavia, Italy, February 2013 Lorenzo Magnani

Contents

Part I Models, Mental Models, Representations

Part II Abduction, Problem Solving and Practical Reasoning

Part III Historical, Epistemological, and Technological Issues

Part I
Models, Mental Models, Representations

Some Ways of Thinking

Barbara Tversky

Abstract When thoughts overwhelm the mind, the mind puts them into the world. Talk, gesture, diagram, and sketch help not only to augment memory and information processing but also to structure thought and promote inferences and discovery. Research from many domains will be presented that highlights the similarities, differences, and special characteristics of each of these tools of thought.

1 Some Ways of Thinking

How do we think? One of those questions that elicits shoulder-shrugs. There are the simple knee-jerk answers: with our brains, or, in words. But there's more to thinking than that. Here I'd like to show, and I'm by no means the first, that we have other ways of thinking. That we take our thoughts out of our minds and put them into the world. Of course we do that every time we talk. But when we talk, we don't just use words, we use the prosody in our words, a bit of which can negate what the words seem to say. When we talk, we use our bodies, our faces, our hands. We use things in the world, pointing to them, arranging them, manipulating them. We use proxies for things in the world, looking toward or pointing to empty places that represent them, where they might have been. Similarly for thinking, which, after all, is communicating with our selves. We think with our hands and our faces and our bodies. We think with the marks and the arrangements of marks we make on paper and the things and arrangements of

B. Tversky (✉)
Columbia Teachers College, New York, USA
e-mail: btversky@stanford.edu

B. Tversky
Stanford University, Stanford, USA

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_1, © Springer-Verlag Berlin Heidelberg 2014

things in the world. You might counter, but all that goes through the brain. Of course, nearly everything goes through the brain. Eating goes through the brain, from the biting and chewing onward. As does walking. Nevertheless, we don't say we eat or walk with the brain.

Perhaps some experiments, thought experiments as well as laboratory ones, will make the case. One of the many reasons for putting our thoughts into the world is the limitation of memory. We make to-do lists, set buzzers and timers, put the shoes that need new soles by the door to remember to take them to the cobbler. In this we are in good company with the rulers of empires who inscribed their accomplishments in stone, often in depictions, not for themselves, but so that others would learn and never forget. Another reason is limits of information processing. Imagine computing the square root of a 4-digit number without paper and pencil. Imagine counting without pointing, and even moving the objects to be counted when their number gets large. When prevented from counting with our fingers, we count with our heads, or with our eyes (e. g., [1]. Sometimes that knowledge is embedded in the actions that produce it. If you're a touch typist, tell me where the keys for "c" or "i" are without moving your fingers. Touch typists typically can't do that without moving their fingers. The brain needs the actions of the fingers to find where the keys are in space.

Thinking with Paper. Putting ideas on paper is common practice for artists, designers, architects, mathematicians, scientists, and ordinary people. Designers refer to having "conversations" with their sketches (e. g., [2]), going to far as to say that the sketch was trying to tell them something. Here's the gist of the conversation: designers, artists, scientists put something on paper that represents their ideas, usually tentative ones. When they contemplate their own sketches, they may discover things in their sketches that they did not intend, they see new things in their sketches, and make inferences from them. Architects, designers, artists, scientists see new relations and configurations (e. g., [3, 4–6]). The new relations and configurations encourage new interpretations. This process–sketching, inspecting, discovering, re-sketching–creates a virtual cycle, a creative one that produces new ideas. A detailed analysis of one experienced architect as he designed a museum revealed that when he reorganized the elements of his sketch, when he saw new configurations, he was more likely to get new ideas, leading to new designs, expressed in new sketches [7]. Expertise matters. In designing a museum, novice architects made many perceptual inferences from their sketches, like noticing sharp angles or finding patterns but experienced architects made more functional inferences from their sketches, for example, inferences about view lines or traffic patterns or changes in light [4]. Intriguingly, the perceptual inferences typically depended on examining the given overhead viewpoint of the museum, but the functional inferences often depending on imagining a different viewpoint, from within the depicted environment.

To further understand the conversation with sketches, we brought the task into the laboratory, borrowing a paradigm of Howard-Jones [8]. We presented ambiguous sketches, those in Fig. 1 below, to participants trial after trial, asking them to produce a new interpretation of the sketch each time they viewed it.

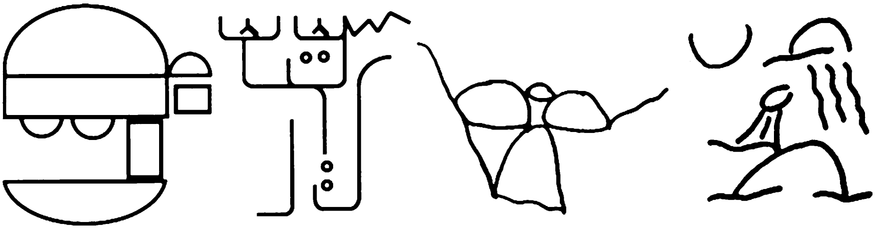


Fig. 1 Four sketches used in experiments of Suwa and Tversky

In various experiments, we presented the sketches to ordinary people, to designers, to architects [9]. In some experiments, we asked people to complete two measures of spatial ability, embedded figures and mental rotation, and a measure of associative thinking, the remote associates task. We counted the number of new interpretations each participant produced for each sketch, and then asked participants how they generated new interpretations.

What did we find? We found that professional designers and architects surpassed untrained participants in numbers of new interpretations of the sketches. We found that those who reported focusing on different parts or reorganizing the parts of the sketch produced more than twice as many new interpretations as those who didn't report that strategy [7]. We found that those who were adept at finding simple figures like triangles and rectangles embedded in complex ones generated more new interpretations than those who were not adept at finding hidden figures [9]. We had already shown that finding new interpretations benefits from perceptual reorganization of complex visual stimuli. We found that mental rotation, a spatial skill requiring imagining objects at different orientations, was unrelated to number of interpretations. We found that those good at producing remote integrative associations, an index of fluid thinking, generated more new interpretations than those who had difficulties in that task. We found that the perceptual reorganization skill was uncorrelated with the associative thinking skill; they were independent of each other Producing interpretations of ambiguous sketches, then, relies on a perceptual skill, generating new organizations and patterns, and a thinking skill, producing a rich web of meaningful associations to those patterns. We proposed that people adept at interpreting and reinterpreting ambiguous sketches rely on a strategy that combines these skills, a strategy we called Constructive Perception, the deliberate and active use of perception in the service of innovative thinking. And we wondered if analogous processes underlie innovative thinking in other domains.

Drawing is used in other domains to sketch out ideas. One such domain is art. Although laypeople may view drawing as transferring what is seen to paper, artists who draw see that process as far more complex (e. g., [3]). They report that drawing is a safe way to explore, that they deliberately get themselves into trouble to break habit and to see if they can find a way out. Like architects and designers, they make unintended discoveries, seeing new things in their own drawings.

Drawing is commonly used in exploration in science as well, beyond constructing clear and simple diagrams for communicating. Scientists look at data every which way to explore, to discover patterns and phenomena. They sketch out explanations to "get the whole picture," to test for completeness and coherence. They sketch to work through problems as well as to communicate known phenomena (e. g., Gooding, Bechtel). Would constructing visual explanations help students as well? Bobek and I presented junior high students with lessons either in the workings of a bicycle pump or in chemical bonding [10]. In both cases, students' knowledge was assessed immediately after the lesson. Then students in each experiment were divided into two groups. One group was asked to provide the typical verbal explanation of the bicycle pump or chemical bonding. The other group was asked to provide visual explanations of the same systems. After completing the explanations, student knowledge was tested again. Importantly, students improved on the second knowledge test simple from constructing explanations, without any intervening teaching. Impressively, students who created visual explanations improved more than those who provided verbal explanations. Examination of their explanations provides clues to the superiority of the visual explanations. They contained more information than the verbal explanations about the structures of the systems. This is to be expected because visual explanations map physical structure in space to the space of a page, a natural mapping, and one of the noted advantages of diagrams. More impressively, the visual explanations contained more information about function, about the operations and causality of the systems. Behavior and causality are harder to depict, and easier to convey in language, yet they were more frequently included in the visual explanations. The visual explanations used language as well as diagrams; the verbal explanations did not add diagrams. But the visual explanations went farther. Just as for artists and scientists, visual explanations provide a check for completeness and a check for coherence: are all the necessary parts there? Do they work together to produce the expected outcomes? The diagrams in the visual explanations provide a natural platform for inferring behavior, process, and causality from structure. Purely symbolic language, with no natural mapping from actual structure to the space of a page does not provide a natural, user-friendly platform for inference and thought.

Thinking with the Body. But what if we don't have paper? Until the twentieth century, paper was rare, and even now we don't have it always with us. What we have with us is our hands, our heads, our bodies, all of which are expressive. Architects, when blindfolded while designing, gesture profusely [11]. Thought has been viewed internalized action; gestures are actions, and can reexternalize thinking [12, 13]. Gestures, like thinking, can pull things together or apart, group things into categories and hierarchies and patterns, turn things upside down or inside out, repeat and delete, arrange and rearrange, and more. Gestures can do even more than reenact thought (e.g., [14, 15]). They can "paint" a scene or an object, iconic gestures. When describing environments, participants use integrated sequences of gestures to place items in their relative locations and to depict features of the items (e. g., [16]). If gestures externalize thought, then perhaps they can facilitate thought. Gestures do appear to offload working memory [17], but they appear to contribute to thinking in more refined and specific ways. Participants alone in a room solving spatial problems gesture when the problems exceed working memory capacity [18]. Their gestures reflect the structures of the problems they are attempting to solve. For a problem about a row of two groups of three glasses each, their gestures were primarily horizontal, corresponding to the two groups of six glasses; to solve a problem about a rising tide and a ladder, they made vertical gestures, corresponding to the rungs of the ladder. Those who gestured were more likely to find a solution to the problem than those who didn't gesture.

If gestures reenact thinking and can facilitate thinking, then gestures that are compatible with the thinking should be more effective than incongruent gestures. Computers with mouse or especially multi-touch interfaces provide an excellent opportunity to test this hypothesis, as well as to teach. Early school-age children from a low SES school were given a series of math problems to solve [19]. Some were discrete problems, notably, addition. Others were continuous problems, finding a specific number on a number line from 0 to 100. The tasks were matched with congruent or incongruent gestures. The congruent discrete gesture for addition was tapping and the congruent continuous gesture for number line estimation was sliding a marker. Children performed better when the gestural actions required by the interface were congruent with the mental actions needed to solve the problems.

Spraction. Both thinking with paper and thinking with the body are thinking with space, and more specifically with actions in space, the actions that create the sketches or the actions of the gestures themselves. The actions in space create simple patterns that abstract and crystallize the thought, a process integrating space, action, and abstraction that I've called *spraction*. We organize our cabinets and closets by putting like things–cups and plates and silverware; socks and sweaters and underwear–together in rows and columns, categories and hierarchies of categories. We distribute plates and silverware and glasses on tables in one-to-one correspondences, place books on shelves and houses on streets in orders. We design the world and the designs carry abstractions. The actions that design space are incorporated into gestures and their patterns into diagrams [20].

I began with an ancient unanswerable question, how do we think? I've tried to show that we think not just in words, but in actions and in the spaces those actions create. I'll end with an example, from a fellow cognitive scientist, Mark Wexler. He used to be a physicist, and worked with Feynman diagrams, elegant simple diagrams of lines straight and squiggly representing particles. Some have double lines, pairs of lines. One day, thinking about a problem, Mark wondered what would happen—he had a hunch–if he picked up the pair like a rubber band and twisted it. It worked. The thinking? Manipulating the diagram with gesture.

Acknowledgments I am indebted to my collaborators Masaki Suwa, Angela Kessell, Juliet Chou, Eliza Bobek, Andrea Kantrowitz, Azedeh Jamalian, Ayelet Segal, Valeria Giardino, and John Gero and to National Science Foundation HHC 0905417, IIS-0725223, IIS-0855995, and REC 0440103.

References

1. Carlson, R.A., Avraamides, M.N., Cary, M., Strasberg, S.: What do the hands externalize in simple arithmetic? J. Exp. Psychol. Learn. Mem. Cogn. **33**(4), 747–756 (2007)
2. Schon, D.A.: The Reflective Practitioner. Harper Collins, New York (1983)
3. Kantrowitz, A.: Drawn to discover: a cognitive perspective. Tracey. www.lboro.ac.uk/departments/sota/tracey/ (2012)
4. Suwa, M., Tversky, B.: What architects and students perceive in their sketches: a protocol analysis. Des. Stud. **18**, 385–403 (1997)
5. Tversky, B., Chou, J.Y.: Creativity: depth and breadth. In: Nagai, Y. (ed.) Design Creativity. Springer, Dordrecht (2010)
6. Tversky, B., Suwa, M.: Thinking with sketches. In: Markman, A. (ed.) Tools for Innovation. Oxford University Press, Oxford (2009)
7. Suwa, M., Tversky, B., Gero, J., Purcell, T.: Seeing into sketches: regrouping parts encourages new interpretations. In Gero, J. S., Tversky, B, Purcell, T. (eds.) Visual and Spatial Reasoning in Design, pp. 207–219. Key Centre of Design Computing and Cognition, Sydney, (2001)
8. Howard-Jones, P.A.: The variation of ideational productivity over short timescales and the influence of an instructional strategy to defocus attention. Proceedings of Twentieth Annual Meeting of the Cognitive Science Society, Lawrence Erlbaum Associates, Hillsdale (1998)
9. Suwa, M., Tversky, B.: Constructive perception: a skill for coordinating perception and conception. In Proceedings of the Cognitive Science Society Meetings (2003)
10. Bobek, E.: Visualizing the invisible: generating explanations of scientific phenomena. Ph. D. Dissertation, Columbia Teachers College (2012)
11. Bilda, Z., Gero, J.S.: Analysis of a blindfolded architect's design session. In: Gero, J.S., Tversky, B., Knight, T. (eds.) Visual and Spatial Reasoning in Design III, pp. 121–136. Key Centre of Design Computing and Cognition, University of Sydney, Sydney (2004)
12. Jamalian, A., Tversky, B.: Gestures alter thinking about time. In: Miyake, N., Peebles, D., Cooper, R.P.(eds.) Proceedings of the 34th Annual Conference of the Cognitive Science Society, pp. 551–557. Cognitive Science Society, Austin TX (2012)
13. Kang, S., Tversky, B., Black, J.B.: From hands to minds: How gestures promote action understanding. In: Miyake, N., Peebles, D., Cooper, R.P. (eds.) Proceedings of the 34th Annual Conference of the Cognitive Science Society, pp. 551–557. Cognitive Science Society, Austin TX (2012)
14. Goldin-Meadow, S.: How Our Hands Help Us Think. Harvard University Press, Cambridge (2003)
15. McNeill, D.: Hand and Mind. University of Chicago Press, Chicago (1992)
16. Emmorey, K., Tversky, B., Taylor, H.: Using space to describe space: perspective in speech, sign, and gesture. Spatial Cogn. Comput. **2**, 157–180 (2000)
17. Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., Wagner, S.: Explaining math: gesturing lightens the load. Psychol. Sci. **12**(6), 516–522 (2001)
18. Kessell, A.M., Tversky, B.: Using gestures and diagrams to think and talk about insight problems. Proceedings of the Meetings of the Cognitive Science Society (2006)
19. Segal, A., Tversky, B., Black, J.: Conceptually congruent gestures promote thought (2012)
20. Tversky, B.: Visualizations of thought. Topics Cogn Sci **3**, 499–535 (2011)

Against Fictionalism

John Woods

Abstract Characteristic of model based science is its attachment to idealizations and abstractions. Idealizations are expressed by statements known to be false. Abstractions are suppressors of what is known to be true. Idealizations over-represent empirical phenomena. Abstractions under-represent them. In a sense, idealization and abstractions are one another's duals. Either way, they are purposeful distortions of phenomena on the ground. Sometimes phenomena on the ground approximate to what their idealizations say of them. Sometimes nothing in nature approaches them in any finite degree. So wide is this gulf between reality and idealization that Nancy Cartwright was moved to say of them that they are "pure fictions".

Fictionalism in the philosophy of science endorses Cartwright's attribution, and occasions an obvious trio of questions. One is whether the fictionality of non-approximating idealizations is a load-bearing contribution to the semantics and epistemology of science. If it is, a second question is whether we know the satisfaction conditions for "is fictional" in virtue of which this is so? The third is whether those satisfaction conditions might profitably be sought in adaptations of what is currently known about the semantics and epistemology of literary fiction.

In this paper I present considerations which, we may come to think, offer these questions scant promise of affirmative answers.

J. Woods (✉)
Department of Philosophy, University of British Columbia, Vancouver, Canada
e-mail: john.woods@ubc.ca

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_2, © Springer-Verlag Berlin Heidelberg 2014

1 Magnetic Pull

Like the month of June, philosophy's attention to fiction is bustin' out all over.[1] What, nearly forty years ago, was a rump movement in the philosophy of language[2] is now the subject of a bustling research programme. The early work on fiction hovered at the intersection of the philosophy of language and analytical aesthetics. Today's range is broader. The concept of fiction is invoked in virtually all branches of philosophy—in the philosophy of science and mathematics; in metaphysics and epistemology; and in ethics and the philosophy of law.[3] The scale and intensity of these developments is striking. Who among those few writing about fiction in the late 1960s and early 1970s could have forseen the decision in 2010 of a major publisher to launch its series on Basic Philosophical Concepts with a book on fiction?[4]

Impressive as it assuredly is, fictionalism's contemporary reach is too much to do justice to in the space available to me here. So in keeping with the conference theme, I shall focus most of my attention on the philosophy of model based science, with a special emphasis on theories that are descriptively intended and designed for experimental test. The thesis that I want to advance is that the fictionalist project for science is a misbegotten one, and ought to be abandoned. My reason for thinking so is that the importation of fictionality into a philosophical theory of science does more harm that good with regard to any end that it might have been intended to achieve. Before presenting a case for this negative proposal, some preliminary matters will require our attention. Let's turn to them now.

When a philosopher invokes fictions for model based science, two questions arise straightaway: *What* is being invoked? And *why* invoke it? The what-question asks for an account of what it is to be a fiction. The why-question asks for the philosophical objectives that fictions are intended to advance, and invites reflection on what it is about them that enables those ends to be achieved. It also asks whether and how fictions *add value* to the theories that call them into play.

Conceived of as a project in the philosophy of language, a theory of fiction develops accounts of inference, truth and reference for fictional discourse,[5] an enterprise which in a suitably flexible sense of the word formulates a *logic of*

[1] With a tip of the hat to Richard Rodgers and Oscar Hammerstein II's "June is bustin' out all over", from the Broadway musical *Carousel*, 1945.

[2] See, for example, Woods [1, 2], Kripke [3], Walton [4] and Howell [5]. The journal *Literary Semantics* was established in 1972 by Trevor Eaton, and Eaton's book of the same title appeared in [6].

[3] Representative coverage is furnished by Woods, editor, [7]. For the philosophy of science and mathematics, see also Suarez, editor, [8], and for mathematics Bonevac, [9]. See also Magnani's [10].

[4] See the note just above.

[5] In mainstream approaches to the semantics of natural language, this order is typically reversed—⟨reference, truth, inference⟩. Reasons for the switch in fictional contexts are laid out in Woods and an Isenberg [11].

fiction.[6] It was widely accepted—and still is—that a satisfactory logic of fiction would, among other things, furnish satisfaction conditions for the predicate "is fictional" and its cognates: "in fiction", "fictionally", "it is fictional that", and their like. In so doing, it would fix their respective extensions: referents, objects, characters, people, events, sentences, narratives, truths, consequences, inferences, and so on. The very idea of a logic of fiction is itself something of an abstraction, instantiated in actual practice by different and often rival approaches, reflecting in turn a considerable variation in semantic assumptions and in the use or avoidance of formal methods.[7] In any event, a logic of fiction is not a symbolic logic. It is not a theory of inference, truth and reference for semi-interpreted formal languages. It is not a logistic system in Church's sense. A logic of fiction is a semantic theory of fictional discourse in natural languages. It is a literary semantics.

No one would think that full coverage of the issues that interest philosophers of even non-aesthetic stripe would be given by this triple of theories I am calling a logic of fiction. A philosopher of mind might be drawn to fictions by an interest in the creative dynamics of story-making, or by the affective etiology of a weeping reader's response to the story that makes him cry.[8] But for my purposes here a certain primacy redounds to the triple, occasioned by the interest that fiction has come to have for the philosophy of science.

Philosophers of science seek clarifications of concepts which strike them as in need of it—the concept of law, for example, or causal explanation, or natural kind. But also high on their agenda are theories of truth for the sentences of a scientific theory, confirmation theories for the theory itself, and theories of inference for the intratheoretic linkages of the theory's sentences—in particular the tie between the theory and its observational test-sentences. Also of interest are the implications of these arrangements for the question of scientific knowledge, and their aggregated impact on ontological commitment and the character of the syntactic vehicles that convey it. It would not be far wrong to take these elements as setting a large and central part of the agenda for the philosophical investigation of science. Of course, dominant though they clearly are, one should not think that these elements exhaust a philosopher's interest in science. A philosopher might be puzzled by the processes of scientific creativity or the influence of societal considerations on scientific research. Even so, when I speak here of a philosophy of science, I will mean, unless otherwise indicated, the advancement of the elements of this core agenda.

I am now in a position to offer an early proposal about scientific fictionalism.

> *The magnetic pull thesis*: When a philosopher of model based science plays the fictional card, fictionalism should be the view that, *in the relevant respects*, the logic of fiction will

[6] As far as I know, the term "logic of fiction", originates with Cohen in [12].

[7] There is not a single equation to be found in, say, Walton's [13]. On the other hand, Parsons' [14] displays a liberal sprinkling of them, as does even more so Jacquette's [15].

[8] Concerning which, see again Walton [4], and Woods and Isenberg [11].

> exercise a magnetic pull on the philosophy of science, and will do so in ways that make for a philosophically more satisfactory theory than would otherwise have been the case.

This tells us something interesting about models (in one of the myriad senses of that word). If Xs are modelled as Ys, then a theory of Ys will exert a magnetic pull on a theory of Xs. The model will pull the modelled into—or in the direction of—its own conceptual space.

In our attributions of fictionality to model based science, there is I think little to be said for numerical identity. An abstractly conceived scientific law is not in any literal sense a fiction. The population-genetic assertion that populations are infinitely large is not strictly a truth of fiction. Less implausible is the idea that when these attributions are made, they are made with a modeller's intent, that is, with the expectation that giving to infinite populations the pull of fictions would redound to the benefit of a philosopher's interest in population genetics.

I find the notion of magnetic pull adumbrated in some well-known words of Nancy Cartwright:

> A model is a work of fiction. Some properties ascribed to objects in the model will be genuine properties of the objects modelled, but others will be properties of convenience. The term 'properties of convenience' was suggested by H. P. Grice, and it is apt. Some of the properties and relations in a model will be real properties, in the sense that other objects in other situations might genuinely have them. But they are introduced into this model as a convenience, *to bring the objects modelled into the range of the* [*modelling*] *theory.*[9]

Cartwright's notion of bringing objects into the range of the theory that models them is nearly enough my notion of magnetic pull. A good example of this is Bayesian epistemology, as reflected in some observations by Hartmann and Sprenger:

> Bayesian epistemology [in contrast to analytically intuitive epistemology], draws much of its power from the mathematical machinery of probability theory, which starts with *mathematical intuition*. The construction of Bayesian models is much triggered by what is mathematically elegant and feasible … The mathematics takes on a life of its own (to adopt a phrase due to Hacking), and the comparison with intuitive examples comes only *after* the Bayesian account is given.[10]

In the notion of pull we find the suggestion of conceptual change. It is not always recognized the extent to which a philosopher's attention to a concept of interest involves a degree of tampering—of tampering, as we might say, for the concept's own good. Even the straightforward clarifications so routinely sought by philosophers more often than not move beyond the exchange of synonyms into the more venturesome territory of explication and, more aggressivily still, rational reconstruction. A question of standing interest for philosophers is the extent to which such transformations leave the original concept recognizably present in the

[9] Cartwright [16], p. 156. Emphasis added. In her text, "[modelling]" is "mathematical." But Cartwright's point is not restricted to mathematical modelling.

[10] Bernecker and Pritchard, editors, [17], p. 629. Emphases in the original.

rethinking of it. It is a question which calls in doubt whether a principled distinction exists between analyzing an old concept and synthesizing a new one. Kant is good on this distinction. Analysis, he says, is the business of philosophy. It is the business of making concepts clear. Synthesis, on the other hand, is the business of mathematics. It is the business of making clear concepts.[11] The magnetic pull thesis attracts these same questions. But the main thing to emphasize here is that fictionalists are of the view—or should be—that our understanding of highly idealized model based science will be improved by reconceptualizing the relevant features, by modelling them, by thinking of them in ways that will add value to a philosophically tenable appreciation of them.

It is prudent to harbour a healthy respect for this difference between an abstractly idealized scientific law and anything a fiction could realistically be taken to be. When a philosopher of science calls upon fictions to do some heavy lifting in his work, he is engaging a "perspective external to the practices" of science.[12] It is easy enough to appreciate the misgivings which so alien a presence might stir in the breast of a cautious philosopher. But there is also the point that, notwithstanding the brisk developments of late, when compared to the philosophy of model based science the logic of fiction is something of a rookie—not only a more recent development but also a good deal less centrally situated in the mainstream of technical philosophy. Even so, we shouldn't over-do this juniority. While there is plenty of occasion to wonder whether any of the going logics of fiction have achieved maturity enough to exercise the drag envisaged by the magnetic pull thesis, this needn't be on account of the recency of these logics. It is generally agreed that the logic of *Principia Mathematica* turned out not to model the arithmetic of the natural numbers. This was not because the logic of *Principia* was new. It was because it was the wrong model. Newness is not the problem with logics of fiction. But callowness might well be.[13]

All this talk of models is dizzying. There is more ambiguity in the word "model" than is perhaps quite good for it. Sometimes a model of something is anything that counts as a simulation of it or, more broadly, any way of thinking it as being, or being like, including ways it could not possibly be. Closer to our current concerns is the threefold ambiguity in which models are structures, or are sentences holding in those structures, or are entities—the things those sentences quantify over. Fictionalism reproduces this trio, yielding up structure-fictionalism, sentence-fictionalism and entity-fictionalism. For our purposes there is no need for trinitarian finickiness. Context will be our guide.

The comparative newness and rawness of the logic of fiction, and of the fictionalisms to which it has given rise, shouldn't distract us from a recognition of fictionalism's ancient lineage, nominalism. When a philosopher is a nominalist about

[11] See Kant [18]. Almost the identical distinction, albeit without mention of Kant, is to be found in Russell's [19], pp. xv–xvi, 15, 27, 112 and 114; originally published in 1903.

[12] In the words of Fine [20], p. 120.

[13] A substantial survey can be found in Woods [21].

numbers he will think that numbers exist in name only. If, in turn, he is a nominalist about fictionality, he will think that numbers are fictions in name only. In that case, his would be a second-order nominalism asserting that it is a fiction in name only that numbers exist in name only. Over the ages, a standing problem for nominalism has been to find for the in-name-only qualification a nontrivial interpretation. It is likewise a problem inherited by the rookie fictionalisms of the present day.

If infinite falsehoods were indeed fictions in name only, the required logic of fiction for model based science would itself have to be a logic of fictions in name only. This introduces a complication. The complication is that it is not clear whether a good logic of fiction would be needed to exert a magnetic pull on the desired logic of fictions in name only. Perhaps it is not too much to assume an affirmative answer. If so, a logic of fiction would have to precede a logic of fiction-in-name-only.

This is not perhaps a welcome complication. Perhaps it makes of fictionalism a slipprier problem than we would have liked it to be. I don't doubt the accuracy of the magnetic pull thesis. If fictionalism is true for science, if fictions are fit for honest work there, a logic of fiction will have to be brought into play, and with it the distinction between a logic of fiction and a logic of fiction-in-name-only. Suppose, however, that fictionalism is not true. Might it not be possible to show this without significant investment in the linkages between logics of fictions and logics of fictions-in-name-only? I will come back to this in Sect. 5.

2 Infinitely False Idealizations

The magnetic pull thesis is relativized to "relevant respects". What would those respects be? Here again is Cartwright:

> A model is a work of fiction … There are the obvious idealizations of physics—infinite potentials, zero time correlations, perfectly rigid rods, and frictionless planes. But it would be a mistake to think entirely in terms of idealization, of properties which we conceive as limiting cases, to which we can approach closer and closer in reality. For some properties are not even approached in reality.[14]

Seeing the importance of this feature of them, Cartwright goes on to say that they are *pure fictions.*[15] They are utter falsehoods utterly on purpose.

It is important to note the cardinality implications of these non-approximating idealizations. Consider two further examples.[16] In population-genetic models of

[14] Cartwright [16], p. 153.

[15] Also important for scientific models are abstractions, whose principal alethic significance is the suppression of what is true on the ground. For further discussion, see Woods and Rosales [22].

[16] Concerning which see also Godfrey-Smith [23]: "Scientists, whose business is understanding the empirical world, often spend their time considering things that are known not to be parts of that world. Standard examples are ideal gases and frictionless planes. Examples also include infinitely large populations in biology, neural networks which learn using biologically unrealistic

natural selection, populations are infinitely large. In neoclassical economics, utilities are infinitely divisible. In the first case, the largest possible and smallest possible actual populations are equally close to the ideal; they both fall infinitely short of it. Similarly, both the smallest and largest numbers by which an actual utility—pleasure, for example—may be divided fall equally close to its ideal; again, they both fall infinitely short of it. To give to these idealizations the name that is due them—"infinitely truth-nonapproximating falsehoods"—would be accurate but stylistically inelegant. I propose a less ponderous description. They are "infinitely remote idealizations". They are falsehoods without a friend in the world.[17] Recognition of the importance of infinitely remote idealizations is widely evident in philosophy of science.[18] Their treatment as fictions is a minority position, albeit one of growing strength in recent years.

As things have so far evolved the magnetic pull thesis hasn't had much sway in fictionalist approaches to science.[19] There is a ready explanation of this. If with respect to a theory's infinitely remote idealizations fictionalism is the doctrine that a logic of fiction will exert a philosophically instructive pull on the logic of science, the structure of the modelling connection needs to be exposed, and its putative philosophical payoff has to be tethered in some disciplined way to features of that structural tie. It would be hard to overstate how far short is fictionalism's present state of play from meeting these requirements.

When you synthesize a new concept into being, you make something up. You make something up in ways that make for the truth of the sentences about its instantiations. You make those sentences true of them, but you make them false of everything in the world. Literary fictions are like this too. At least, they are somewhat similar. When Doyle made Holmes up, he did so in ways that made various things true of Holmes, for example, that he shared rooms with Watson at

(Footnote 16 continued)
rules, and the wholly rational and self-interested agents of various social-scientific models …. A natural first description of these things is as fictions, creatures of the imagination." (p. 101).

[17] It might be thought that these infinite gaps could be made subject to variable shrinkage by the devices of probability. A statement having a probability of 0.8 is one with a higher probability than a statement whose value is 0.6. This is so, but not on point. The highest probability is 1.0. Perhaps we could say that statements having this value are ideally probable. At least, if we did say this, people would know what we meant. But 0.6, 0.8 and 1.0 are real numbers. Real numbers are everywhere dense. No real number (save for self) is any closer to 1.0 than any other. The relation of having a higher probability than is not a matter of having a value that lies closer than the alternative to the ideal probability. Of course 8 is a number that lies closer to 10 than 6 does. But these are natural numbers. Other than 1 and 0 numbers on the natural line are not probabilities.

[18] Batterman, among others, writes astutely about the philosophical questions raised by the ineliminable presence of unapproachable idealizations in theoretical science, but with no mention of the idea that they are fictions. See, for example, his [24, 25]. Other sceptics of note are Teller [26] and Giere [27], both in Suarez [8].

[19] This is starting to change. Two important exceptions are, Suárez [28] and Frigg [29], both in Woods, *Fictions and Models.* In these approaches, the borrowed treatment of fictions is the so-called pretense theory of Kendall Walton. See again his [13].

221B Baker Street. There are truths to which nothing in the world approaches in any finitely realizable degree. Of course, some real-world people share rooms at some real-world address, but this is not the comparison we seek. Nothing that is remotely possible of realization in the real world approximates in any finite degree two unreal people sharing rooms in a real-life city. We would seem to have it, then, that the truths of literature are infinitely remote from the truths of the world and they are truths that their tellers make up. So isn't it true to say that these makings-up of science are fictions?

The short answer is yes. That is, yes up to a point; indeed, up to the two points of similarity we've taken notice of just now. With this resemblance in mind, when you characterize an infinitely remote idealization as a fiction, you are attributing to it two of the characteristics that everyone already knows such idealizations to possess. What you assert is a two-part similarity. The similarity clearly exists, but noticing it exercises no magnetic pull on the concept of the infinitely remote idealization.

3 Outsourcing

Fiction is a borrowed concept in any philosophical theory of science that invokes it. The borrowed concept is a concept "external" to the conceptual space of the borrowing theory. Concept-borrowing theories reflect a distinction between a fiction's work-producing status and what we might call its *façon de parler* status. For philosophers such as Arthur Fine, "fictionalism" is just another name for antirealism,[20] for the view that a sentence such as "Numbers are fictions" is only a lexical variation of "Numbers aren't real". In this usage, numbers are fictions in a manner of speaking. It is not hard to see why *façon de parler* fictionalism would not be of much interest to the philosopher of science.[21] Although the *façon de parler* variant is, rather more than not, fictionalism's standard form, it is rarely acknowledge as such. Remarks by Alisa Bokulich provide an instructive example of this fact. She writes:

[20] Fine writes: "Over the last few years the realism-antirealism arguments … and a somewhat larger number of epithets …. When an especially derisive antirealist label is wanted, one can fall back on the term "fictionalist", coupled with a dismissive reference to Vaihinger and his ridiculous philosophy of 'As If'". ("Fictionalism", *Midwest Studies in Philosophy,* 18 (1993), 1–18, p. 1.) Fine's use of "ridiculous" is a matter of mention rather than use. Fine is no Vaihingerian, but he is far from thinking that *The Philosophy of 'As If'* is ridiculous.

[21] This is not to say that *façon de parler* fictionalism is inherently antirealist. Fictional realists hold that Holmes is a real thing, albeit not in the way that we ourselves are. Realistically inclined idealizers claim that infinite populations are real, albeit not in the way that the population of London is. If the attribution of fictionality to those idealizations is just another way of saying that they are real, but not in the way that the population of London is, the attributor is a *façon de parler* fictionalist.

> As we know well today, however, these Bohr orbits are *fictions*—according to quantum mechanics the electron in an atom does not follow a classical trajectory in a stationary state and is better described as a cloud of probability density around the nucleus. I want to defend the view that, *being a fiction*, Bohr's model of the atom does in fact explain the spectrum.[22]

We have it, then, that orbits are fictions because they are not truths of QM. They are idealizations without a friend in the QM-world. They are infinitely unapproachable falsehoods in Cartwright's sense and infinitely remote falsehoods in mine. Granted that such idealizations have genuine explanatory force in QM, it is easy to see that the Bohr model is a well-motivated contrivance. But what, beyond calling them infinitely unreasonable idealizations, is added by also calling them fictions? What is it about "being a fiction" that renders an unrealizable falsehood capable of explanatory force that "being an unrealizable idealization" lacks"? Finding that this question is not posed in "How scientific models explain", I think that we must conclude that Bokulich's fictionalism is also of the *façon de parler* kind.

Let's say no more of *façon de parler* fictionalism. The fictions we want for science will have a more load-bearing role to play. The magnetic pull thesis requires no less of them. The scientific fictionalist has two broad borrowing options.

> *The homegrown* option: A logic of fiction for a concept-borrowing theory—e.g. the philosophy of population biology—will be *sui generis.* The theory will construct its own purpose-built treatment of fiction.
>
> *The outsourcing option*: A logic of fiction for a concept-borrowing philosophical theory will be a borrowed logic of fiction, suitably adapted. Theories exercising the outsourcing option are thus borrowers of fiction twice over. They are borrowers of the concept and they are borrowers of a logic of it.

If we examine the current literature, we see that when substantive borrowing actually occurs, the outsourcing option is the almost universally exercised one, and that the source of the borrowing is typically some or other theory of literary fictions.[23]

Whichever option we might decide to exercise, an earlier question presses for answers. It is the value-added question. What of importance would be left out of a philosophical account of model based science if it didn't assign a working role to fictions—the why-question—or, having done so, didn't also provide an independent account of what it is to *be* a fiction—the what-question? These questions

[22] Bokulich [30], The emphasis, in the first instance, is hers, and in the second mine.

[23] See again Suárez and Frigg. An exception is Vaihinger's [31], the book arises from Vaihinger's doctoral dissertation of 1877. Meinongean theories, in turn, are adaptations of an antecedently developed metaphysical theory. See again Parsons' *Nonexistent Objects* and Jacquette's *Meinongean Logic.* Also important is Bonevac's home-made mathematical fictionalism, in "Fictionalism", Sects. 7.2–7.8, made especially interesting by the fact that Bonevac is not himself a fictionalist.

apply equally to both options, homegrown and outsourcing. But since outsourcing is our focus here, we should direct the question to it.

4 Neutral Fictionalism

When Xs are modelled as Ys, the pull of a theory of Ys on a theory of Xs can be deflationary, inflationary or neutral. A neutral pull is actually no pull at all. It is pull in name only. If “Y” is just another name for X—or concretely, if “fictional” is just another name for “infinitely remote idealization”—then idealizations aren’t modelled as fictions, and a logic of fiction exerts no pull on the logic of science in respect of them. Fictions in this sense are fictions in a manner of speaking. Again, they are *façon de parler* fictions.

Let us come back to the point that sometimes the attribution of fictionality is intended to call attention to a similarity between some feature of fiction and some aspect of the thing to which fictionality is ascribed; that is, it serves as a simulacrum of it.[24] Everyone knows that fictions are made up by authors. Everyone knows that idealizations are made up by theorists. They are made up by their progenitors in ways that make them stick. Sir Arthur Conan Doyle made it *stick* that in the stories Holmes lived at 221B Baker Street. Population biologists made it *stick* in genetic theories of natural selection that populations are infinitely large. No one believed it when Quine said that theories are free for the thinking up.[25] The truth is that making the makings-up of population biology stick is subject to complex and not easily discerned conditions. Something like this is also true of Doyle, though more gently so. Not even Doyle can make it stick that Holmes lived in Baker Street without finding the contextual conditions that enable it. These would be resemblances worth making remarking on if conditions for making things stick in fiction exercised a magnetic pull on conditions for making things stick in population biology. But there hasn’t to date been a jot of evidence to support the existence of that pull, and lots to support its nonexistence.[26] Until something more convincing comes along, the prudent course is to take similarity fictionalism for what it is. It too is fictionalism without conceptual pull.

We now have the wherewithal to sound a gentle admonition:

Lesson 1: Avoid *façon de parler* fictionalism and like varieties with neutral pull.

It is sometimes said (by physicists) that there are two matters in which biologists exhibit a notable deficiency. They aren’t as adept as they should be at data

[24] Virtually all the going fictionalist accounts of mathematics are *façon de parler* accounts or similarity accounts. See again Bonevac’s [9]. See also Cartwright’s discussion of representation in [16], p. 156.

[25] Quine [32].

[26] See here Woods and Rosales [33].

analysis; and they haven't quite got the hang of model construction. Perhaps this is a bit too partisan a complaint, but let that pass for now. For our purposes it suffices to ask: "Suppose a biologist wanted some instruction in the building of powerful models. Should he enroll himself in a course on creative writing, or should he hire a physics post doc?" Early papers of Patrick Suppes contain helpful admonitions about these and related matters.[27]

5 Unreasonable Effectiveness

Let the genetic model of natural selection be our guide. The model provides that populations are infinitely large. In so doing, it fails (utterly) to tell us anything true about population size. But also in so doing, it helps tell us something true about natural selection in real populations—in populations on the ground. The ensuing theory is an empirical success. It performs well at the empirical checkout counter. It is a theory that gives us a knowledge of how natural selection actually works. The idealization of infinite largeness is no mere heuristic. It plays an ineliminable role in generating the theory's results, the results to which it owes its empirical adequacy. A good deal of what is philosophically interesting about this branch of population biology is that without its infinitely false provision for population size its empirical adequacy would collapse. It is a falsehood that can't be "de-idealized."[28]

On the face of it, this is an epistemically discouraging dependency, nicely described by Eugene Wigner as an "unreasonable effectiveness".[29] The unreasonable effectiveness problem has spawned a sizeable and contentious literature that touches in one way or another on virtually every issue of significance for the epistemology and metaphysics of science, including every way of being a realist and every way of not being one. To the extent possible, I want to avoid these entanglements. They are, to be sure, matters of importance, but they are not the engagements for which my project has been fashioned.

There are two cases to which I want to give particular attention—two problems that seem to arise quite naturally in the contexts presently under consideration. The first is posed by

> *The detonation question*: How widely spread in a theory T is the alethic impact of its ineliminable idealizations? How contagious is the property of infinite falsehood?

The second question is to be found in recent remarks of Nancy Cartwright:

> Although on some theories of fictions, fictional descriptions need not be false to the real world, it is nevertheless characteristic of fictions that they are. This seems to be the central

[27] See Suppes [34, 35].

[28] I borrow the term from McMullin [36]. See also Suarez [28].

[29] Wigner [37].

> feature literary features share in common with the claims of mathematics of metaphysics and of many scientific models. From this starting point Woods' challenge … is to explain how focus on this shared characteristic helps solve the problem in view, which in my case is to understand the use of false models *to generate* true claims about target situations. That, it seems, can be a tall order.[30]

This we might call

> *The grounding question*: A principal task of an empirically intended theory T is the generation of the observation sentences O_i on which its empirical adequacy ultimately depends. Schematize this generation relation as $T \vdash O_i$. Generation is a kind of derivation or grounding relation. But given T's dependency on infinitely false idealizations, derivations in the form $T \vdash O_i$ are unsound. Do we not have it then that T lends no grounding support to the O_i it generates, and to which it owes its empirical adequacy? The grounding issue is a nasty looking problem. When a theory carries a structure in the form $T \vdash O_i$, and the O_i are the observational consequences to which T owes its empirical adequacy, and yet T cannot properly speaking be said to ground its observational consequences, how does it get to be the case that the O_i's truth reflect well on T? Intuitively speaking the O_i reflect well on T because they are results for which T is responsible. How can T take the credit for the O_i if it makes no case for them?[31]

Let P be any problem—these two or any other—putatively occasioned by the unreasonable effectiveness of a model based theory T. Let the invocation of fictionality be intended as a substantive contribution to P's resolution R. My question here is not whether R resolves P but rather, assuming that it did, what would it have been about the invocation of fictionality that made or helped make it so? A second question is more ambitious. Once we started paying attention to what they actually are, might it not turn out that fictions possess a property Q thanks to which, for any R that resolves P, fictions make no contribution to it? If so, the invocation of fictions would add no value to P's resolution. This second question helps box my compass. It gives me a strategic option to consider. It motivates the search for an affirmative answer. It motivates the search for a relevant Q. Let us call this the *Q-strategy*.

Suppose further that a search for the relevant Q could be achieved without having to produce a full-bore logic of fiction. This would be a boon twice-over. The defeat of fictionalism would be both *principled* and *cheap*. Whereupon, a second strategy announces itself.

> *The pre-emption strategy*: To the extent possible, one's search for Q should avoid the need for a full-bore logic of fiction.

Again there is a reason for this caution. Slightly over-simplified, there are more logics of fiction than you can shake a stick at.[32] The more that one's search for Q is wedded to a given logic, the more it imbibes the dissensus that surrounds it. Of

[30] Cartwright [38]. Emphasis added.

[31] The distinction between consequence and ground is a crucial in all case-making contexts, yet the logic of grounding is not as technically well-advanced as one might expect it to be. For recent work, see Fine [39].

[32] See again my "Fictions and their logics", in Jacquette's *Philosophy of Logic.*

course, fictionalism aside, there is also considerable dissensus in the logic of science about how to handle Wigner's problem and others. The fictionalist's goal should be to minimize that dissensus, if not remove it outright. It is remotely possible that throwing a contentious logic of fiction into the mix could turn out well. Perhaps there would be synergies that offered relief in both directions, that calmed the philosophical waters of fiction and science alike. But it is not typically the case that adding dissensus to dissensus is the way to bring about this kind of amity. An indication of the extent and depth of these rivalries can be found below in an appendix.

With these questions now at hand, let me say again what I intend my task to be. Without having to develop a comprehensive logic of fiction (the pre-emption strategy), I will try to find a property of fictions Q in virtue of which, depending on the problem under consideration, one or other of two results is achieved. Either Q condemns the problem to irresolution, or it shows that fictions make no contribution to its solution if indeed it has one (the Q-strategy).

6 Empirical Adequacy

The detonation and grounding questions arise from a worry about how a theory ineliminably tethered to an un-de-idealizeable falsehood could ever manage to achieve empirical adequacy. The technical notion of empirical adequacy has been a central topic in the philosophy of science since 1980 when Bas van Fraassen advanced a detailed account of what we should mean by it.[33] There is wide, if not universal, satisfaction with van Fraassen's characterization, and that is good enough for what I am about here. The intuitive notion is that

> A theory is empirically adequate exactly if what it says about the observable things and events is true—exactly if it 'saves' the phenomena' (p. 12).

More formally,

> To present a theory is to specify a family of structures, its models; and secondly, to specify certain parts of those models (the empirical substructures) as candidates for the direct representation of observable phenomena (p. 64).

Then a theory is empirically adequate "if the structures which can be described in experimental and measurement reports" are isomorphic to the empirical substructures of some model of the theory.

Van Fraassen's further claim that for theories of the sort under review, empirical adequacy is the epistemically best to hope for, that empirical adequacy doesn't confer truth upon their nonobservational sentences. This is van Fraassen's constructive empiricism. In the spirit of wanting to stay out of unnecessary trouble, I accept van Fraassen's construal of empirical adequacy but stand mute on

[33] Van Fraassen [40].

constructive empiricism.[34] I want to have my say about fictionalism in science without having to settle the hash of constructive empiricism.

7 Observational Consequences

The unreasonable effectiveness questions place a not inconsiderable burden on the structure we are schematizing as $T \vdash O_i$. Here, too, we meet with a thicket of difficult and largely unresolved philosophical controversies, indeed three of them. One is a snarl of questions concerning how to interpret the $\vdash$-relation(s). The other two concern the interpretation of $\vdash$'s relata. Here, too, I want to stir clear of these entanglements. This is the right way to proceed. At least it is the right way to proceed if the Q-strategy admits of an execution that permits it. This it would do if we were able to find the sought-after Q without heavy investment in adjacent matters. I will say that such a Q is indeed findable, provided we make some assumptions about empirically intended model based science. One is that it is sometimes empirically adequate. Another is that when this is so it is made so by the success of its observational consequences at the empirical checkout counter.

8 Deflationary Fictionalism

Neutral fictionalism is fictionalism without pull. Deflationary deflationism is something different. The same is true of inflationary fictionalism. A deflationary fictionalism with respect to a theory is a fictionalism that takes something *out* of it. An inflationary fictionalism with respect to a theory is a fictionalism that puts something *into* it. Deflationary theories are well-known to philosophers, even to those who have slight acquaintance, or none at all, with literary semantics. Arguably, the most widely discussed is Russell's doctrine of logical fictions.

Logical fictions are logical constructions by another name. At *Principia Mathematica* 20, Whitehead and Russell discuss the no-class theory of classes, and conclude that classes are logical fictions. A better known example is logicism, the thesis that arithmetic reduces to *PM*'s pure logic, that is, to the fusion of quantification theory and its treatment of set theory. The purported reduction turns on a truth preserving equivalence relation under which the theorems of number theory are re-expressible without relevant loss in the language of pure logic. Thus the natural numbers are logical fictions too.

It is widely believed that logicism was motivated by a determination to slip from the embrace of ontologically licentious entities. In fact this is a Quinean wrinkle. *Principia*'s central motivation for the reduction of arithmetic was to

[34] For telling objections, even so, see Rosen [41].

favour proof over postulation. Without the reduction, some of the statements required by arithmetic would have to be introduced without proof, that is to say, by axiomatic stipulation or by what Quine would later call discursive postulation.[35] According to Russell,

> The method of 'postulating' what we want has many advantages; they are the same as the advantages of theft over honest toil. Let us leave them to others and proceed with our honest toil.[36]

If logicism is true, numbers are eliminable without cost to the truths of arithmetic. These days hardly anyone thinks that logicism *is* true.[37] But, true or not, we have a perfectly manageable concept of fictionality at our disposal. Perhaps the logical fictions of *Principia* didn't do the work intended for them in Whitehead and Russell's epistemology of mathematics. But that needn't have been because of a deficiency in the very idea of them. Nor need it have precluded that concept's utility in different contexts of philosophical enquiry. Why could it not be given consideration for use by the scientific fictionalist?

No. The goal of the logical fictions programme was to preserve the truths of arithmetic in a way that severed their apparent commitment to numbers. This is achieved by an equivalence between the sentences of number theory and sentences of pure logic, which latter carry no appearance of a commitment to numbers. In this way "2 is the only even prime" could be true without anything's being the number two. Clearly this won't work for the infinite falsehoods of model based science. Suppose that for "Populations are infinitely large" there were a truth preserving equivalence with a sentence carrying no apparent commitment to infinite populations. Then we would lose the necessity to recognize infinite populations, but we wouldn't preserve the truth of "Populations are infinitely large". For it is false that populations are infinitely large. When we say that the number two is a logical fiction, we enable ourselves to say that "2 is the only even prime" is true even if there are no numbers. If infinite populations were logical fictions, then "Populations are infinitely large" would be true even if there were no infinite populations. Since "Populations are infinitely large" is not true, infinite populations aren't logical fictions. So, we have a second lesson to propose.

Lesson 2: Avoid logical fictions.

A further source of deflationary difficulty can be found in what Bentham has to say about fictions.[38] Bentham's fictions intersect without present interests in three

[35] Besides, Whitehead and Russell didn't think that classes had anything like a decisive ontological advantage over numbers. (Why else would they advance the no-class theory?).

[36] Russell [42], p. 71. Reprinted 1993.

[37] That is, the original project failed. Attempts to rescue significant parts of logicism have been attempted over the years, some making notable progress. See here Burgess [43].

[38] Bentham [44], It is a matter of note that Ogden is Vaihinger's English translator. Bentham, by the way, should not be confused with his nephew the logician George Bentham. See Bentham [45].

ways, two of which have deflationary significance. Under the encouragement of Odgen and (especially) Quine, Bentham is seen as the holistic precursor of the semantic preference for the sentence over the term, and, in his doctrine of paraphrasis, adumbrator of contextual definitions of the sort that underpin the reductive cleansings of *Principia.*[39]

> [A] fictious entity is an entity to which, through the grammatical forms of the discourse employed in speaking of its existence is ascribed, yet in truth and in reality existence is not meant to be ascribed."[40]

In their use as contextual eliminabilities Benthamite fictions, no less than the logical fictions of the logicists, are not load-bearing in any theory in which they occur. A logical fiction in a theory is something that is not needed for its truths. But, again, if we effected a Benthamist reduction which removed infinite populations from population genetics, we would leave intact the infinite falsity of "Populations are infinitely large". We must conclude, therefore, that a philosopher of science of fictionalist leanings would be wasting his time outsourcing his interest in fictions to Bentham's provisions for them. A fictionalism in the manner of Bentham would, as with Russell's, be a half-hearted deflationism.

9 Whole-Hearted Deflationism

Half-hearted deflationism gets rid of entities but not of the sentences which carry the appearance of commitment to them. This is fine if the sentence was true from the outset and remains true once its apparent referents have been made to go away. Deflation is whole-hearted when the sentences up for treatment are false and yet can be made to go away without collateral damage to theories in which they occur. Half-hearted fictionalist deflation gets rid of entities. Whole-hearted fictionalist deflation gets rid of falsehoods. Let's consider this now.

It is natural to see in the infinitely remote falsehoods of model based science a connection to a family of problems surrounding a theory's theoretical sentences, sentences that lie strictly beyond the reach of the observable. One of philosophy's standing worries is whether theoretical sentences have any epistemically defensible place in empirical science. One solution is to get rid of a theory's theoretical sentences without damage to its observational power—in other words, to defeat the presumption of indispensability. Two of the better-known examples of this approach involve Ramsey sentences and the Craig elimination theorem.[41] Both these contributions raise technical matters requiring a certain adroitness of

[39] See Quine [46], pp. 67–72. Quine writes at p. 69: "It was the recognition of this semantic primacy of sentences that gave us contextual definition. I attribute […] this to Bentham."

[40] Bentham [47].

[41] Ramsey [48]. Craig [49–51]. See also his [52].

exposition and analysis. But we can make do with brief informal sketches that give their gist.

Beginning with Ramsey, suppose that the vocabulary of a theory contains the name "neutrino" and that many of its theorems describe the properties that neutrinos possess. Let Φ be a neutrino sentence ascribing property F. Then a Ramsey sentence with respect to Φ and F arises from Φ by replacement of "neutrino" with an individual variable and application of the existential quantifier. Whereas Φ purported to say that neutrinos have property F, the Ramsey sentence says that property F is possessed by something. Thus, the Ramsey sentence is topic-neutral, whereas Φ is committed to an unobservable entity.

There is a similar basis to Craig's elimination theorem. Suppose we divided the terms of a scientific theory into observational terms *o* and theoretical terms *t*. Then if there exists a logistic system S which formalizes the theory and the theory gives a set of purely *o*-consequences, then there is also a system S^o containing only *o*-terms that gives those same *o*-consequences.

One might think that either way, Ramsey's or Craig's, the problem of falsehood-indispensability for science is now solved, that whole-hearted deflationism is the way to go. If, for example, the falsity of the sentence "Populations are infinitely large" arises from the fact that "population" is a theoretical term purporting to denote a theoretical entity, then the fact that the population genetics can get on with its job without the term, or its putative denotatum, removes the ground for the complaint that the theory is indissolubly wedded to a falsehood. This gives us two alternatives to consider -the falsehood-indispensibility problem is solved in one or both of these ways, or it is not. If not, that is the end of the matter. It remains unsolved even if we call the denotata of *t*-terms fictions. But if the problem is solved, then the problem is *solved.* There is no need to invoke fictions.[42] Fictions would be surplus to need.

We can generalize on such cases. The trouble that contextual elimination gives to the fictionalist is that, although truth preserving, it is not *falsity removing.* Since our present problems pivot on the infinite falsities of model based science, surely the sensible course would be to search for an equivalence which, under the requisite constraints, mapped those falsities to truths. Let L be a sensory language—a language of how things appear. Let M and M* be two models of a scientific theory, M containing the theory's idealizations and abstractions and M* entirely free of such. Then M and M* are elementarily equivalent with respect to L just in case they satisfy the same sentences of L. For the problems currently in view, why

[42] Apart from their potential as indispensable-falsehood solutions, Ramsey and Craig eliminations are faced with internal difficulties. If, for example, we Ramseyized an entire theory, except for its logical particles, then Löwenheim-Skolem considerations would now apply and with them, some would say, the loss of the theory's scientific content. As for Craig's claim that the *o*-consequences are invariant under the transition from S to S^o, there is no effective way of producing these *o*-consequences in the first place without the aid of *t*-terms. So, there is an important sense in which the transformation doesn't cancel *t*-term dependencies.

wouldn't the prudent course be to search for an equivalence with respect to L that deflated a model based theory's M in favour of an equivalent M*?

Of course, there is always the question of whether such an equivalence can be found or convincingly presented. That is a matter of some moment for those who want to replace a theory's load-bearing falsities with sentences not thus stricken. For our purposes, however, it is not a pressing issue. For either such equivalences are convincingly available or they are not. If they are, there is no problem occasioned by a theory's ineliminable falsities, hence no problem to be solved or even influenced by the attribution of fictionality. If, on the other hand, the sought for equivalences aren't convincingly available, the falsities of the theory stay ineliminably in place, and the invocation of fictionality does nothing to deflate them. This gives us a third lesson to consider.

> *Lesson* 3: Whole-hearted deflationary measures for infinitely false idealizations dispossess fictionalism of a coherent rationale.

10 Dyadicizing Truth

Against this it could always be proposed that there is indeed a coherent rationale for fictionalizing the remote falsities of science. Virtually everyone agrees that the sentences of a story are false (or anyhow not true), and yet a good many insist that they are also true. The inconsistency that looms in so saying is disarmed by a plea of ambiguity. "Holmes lived in Baker Street" is false in actuality but true in the story. Indeed it would seem that the central task of a theory of truth for fiction is to sort out the details of this dyadicization of truth—truth in actuality and truth in fiction. Why, it might be wondered, couldn't we exercise this same option for model based science? Why couldn't we find a sensible basis for distinguishing truth in actuality from truth in a model (or truth in a theory)? And why couldn't an account of truth in models be expected to yield to the pull of a theory of truth in fiction? It is an interesting suggestion, raising more questions than there is space for here. Even so, it strikes me that one particular difficulty stands out. It is that no literary semanticist thinks that fictional truth cancels actual falsity, that it deflates it. In making it true in the story that Holmes lived in Baker Street, Doyle never intended to override its actual falsity. Doyle was not trying to add to London's population without, so to speak, benefit of clergy. The same applies to truth in a model. If it is false on the ground that populations are infinitely large, it is a falsity undisturbed by the truth in the model of its negation. Truth in a model doesn't wipe out falsity in the world. Neither does truth in a model cancel the difficulties occasioned by falsity in the world. Accordingly,

> *Lesson* 4: Avoid the truth-dyadicization strategy.

11 Inflationary Fictionalism

Deflationism is an approving response to Ockham's injunction not to multiply entities beyond necessity. Inflationism pulls in the opposite direction. It is an approving response to the admonition to multiply entities as may be needed. Examples abound, not the least of which are the nonconservative extensions of logic and mathematics. Inflationary manoeuvres add to a theory items it previously lacked—a new axiom, a new transformation rule, an abstraction or idealization that alters the theoretical landscape in some significant way, and so on. It can now be appreciated that the magnetic pull thesis embodies a strong approval of inflationism in model based science. If Xs are modelled as Ys then, as the thesis attests, a theory of Ys will exert a magnetic pull on a theory of Xs. If Xs are modelled as fictions, a theory of X should yield to the pull of what a theory of fiction calls for. Fictions, as described in that theory, should be given honest work to do in the theory of Xs. Fictions should inflate the theory of Xs to its advantage. The magnetic tug thesis says that inflating a model based theory of science with fictions, without the guidance of a theory of fictions, is inflation to no good end. It is an inflation that adds no value.

I have already said that fictionalists about science hardly ever pay for their fictional inflations with a theory of fictions, although most of the comparatively few who do try to pay their dues by harnassing the literary theory of Kendall Walton. Whatever the merits of Walton's pretense theory,[43] this is very much the right way of proceeding. That is to say, it is the right way of proceeding if fictionalism itself is the right way to proceed.

But it isn't, as witness now Bentham's treatment of *legal* fictions.

12 Detonation

I said that Bentham's fictions intersect with our present interests in three ways. The first two have to do with the primacy of sentences over terms and the paraphrastic eliminability of terms without damage to truth. The third way is something quite different. Fictions in this third sense arise from the view that legal facts are fictions, for example, the legal fiction that corporations are persons. This, says Bentham, is a fact generated by social policy, in particular, by the desire that corporations be subject to the laws of tort. Legal fictions are made distinctive by virtue of the fact that they are created and given force by the human will and are maintained by society's determination to be governed by them as if they were the real thing.

At first blush, we might well suppose that legal fictions are an attractive inflationary possibility for outsourcing fictionalists. Perhaps their most agreeable

[43] Reservations are advanced in my "Fictions and their logics".

feature is that legal fictions are stipulated into being in accordance with their inventors' overarching interests. In the legal case, those interests are the requirements of justice broadly speaking. In the scientific cases, it will matter whether or not the inclusion of a fiction has the effect of fictionalizing the whole theory in which it has been placed. On the legal side, the answer is in the affirmative. Decisions that consummate legal proceedings issue forth in legal facts (the legal fact that the accused is guilty, the legal fact that damages are owed, and so on). But no one seriously supposes that a legal fact always has a counterpart actual fact.[44]

This is an important feature of legal fictions. Legal fictions are subject to what we might call "the semantic integration property". Legal facts combine with real facts to produce further facts. It is a legal fact, even if not an actual one, that the Acme Bank is a person. It is an actual fact that office-holders of the Acme Bank defrauded its clients. So it is a further fact that the person that the Acme Bank is owes damages to its clients. On Bentham's understanding, there is an additional feature to take note of. In semantic integration contexts, the property of legal facthood is passed on to dependent facts. The fact that the bank owes damages to its clients is a legal fact even if (for metaphysical reasons) it couldn't be an actual fact. Accordingly,

> *Dependency distribution*: A characteristic of Bentham's legal fictions is that the fictionality property is distributed to dependent sentences in semantic integration contexts.

Because legal fictions have the dependency distribution property, they provide an affirmative answer to a detonation question of its own. It is the question of how widely spread is the legal fiction property in semantic integration contexts. Bentham's answer (and I think the right one) is that it is utterly contagious in those contexts. It detonates there. It is striking that the detonation property is not peculiar to legal fictions. It holds of them then not because of their legality but rather because of their fictionality. It is fictional fact that Holmes is a man and a real fact that men have oesophaguses. So it is a fictional fact, not a real one, that Holmes has an oesophagus. It is a fictional fact that Holmes lives in London and an actual fact that London is in England. So it is also a fact that Holmes lived in England, not an actual fact but a fictional one.[45]

The dependency distribution property causes literary fictionality to detonate in semantic integration contexts. If the infinitely remote idealizations of model based science were fictions, they too would have a dependency distribution problem. The property of infinite falsity would denotate in semantic integration contexts. An empirically intended model based theory T is just such a context. False sentences combine with true sentences in ways that instantiate $T \vdash O_i$. If the infinitely remote

[44] This is especially true of criminal cases at common law. Acquittals constitute the legal fact of innocence. But legal innocence significantly outpaces actual innocence. This is deliberate. It arises from a social policy designed to minimize wrongful convictions. Better a false acquittal than a false conviction.

[45] Unless, of course, Doyle provides otherwise. Either way, these are fictional facts, not real ones.

falsehoods ineliminably embedded in T detonated there, the dependent O_i would themselves be forlornly false, and T could not imaginably fulfill the conditions required for empirical adequacy.

This is a setback for the fictionalist project in science. Fictions detonate and infinite falsehoods don't. So infinite falsehoods can't be fictions. This gives us the sought-after property that fulfills our Q-strategy. It does so in a way that also executes the pre-emption strategy. No full-bore logic of fiction is needed to recognize that fictions detonate in semantic integration contexts.

In a way, the detonation question for forlorn falsehoods was a trick question. It is a logical commonplace that, unlike truth, falsity is not preserved under consequence. How surprising can it be, then, that when $T \vdash O_i$ holds, the falsity embedded in T is not passed on to the O_i? The very fact of T's empirical adequacy precludes the detonation of its falsities. It is precisely here that fictionality's explosiveness achieves a grip. Since detonation is not a problem for falsely tinctured Ts, fictions are not required to fix it. Yet if fictions were called into play, they would *create* a denotation problem for T, and would guarantee that it could not be solved. For, again, detonation precludes empirical adequacy.

I take this to be a serious discouragement of the fictionalist programme for science, and it bears in an interesting way on the grounding question. If the O_i of an empirically adequate T are underivable in the absence of T's infinitely remote falsehoods, then T's connection to those O_i cannot be grounding. T cannot be said to have demonstrated those consequences or to have provided a reason that supports them. This is a puzzle. But suppose, now, that fictions were called into play with a view to solving it. Then T wouldn't be empirically adequate. (Fictionality detonates.) The grounding question asks how T can be empirically adequate if it doesn't lend grounding support to the O_i in virtue of which this is so. But if fictions are let loose here, the empirical adequacy of T is lost. The grounding question wouldn't arise.

I don't want to end this section without some mention of Vaihinger. Vaihinger's *The Philosophy of "As If"* is a work of importance whose neglect is something to regret, and whose repair is beyond what I have space for here. Vaihinger's enthusiasm for fictions is striking and, one might think, excessive. It is sometimes hard to see how anything manages not to be a fiction of some or other Vaihingerian sort. There are ten different types of them, some admitting of subtypes. It is not easy to capture what lies in common among so aggressive a variety, but one thing is clear. Vaihinger seems to think—although not in these words—that fictions have the dependency distribution property, that the property of fictionality detonates in semantic integration contexts. Vaihinger is an unapologetic instrumentalist.[46] Theoretical science may serve us well or badly, but a scientific theory is never true even when good.

[46] See again Fine's "Fictionalism" and Bonevac's "Fictionalism", Sect. 3.4.

13 Non-empirically Intended Theories

So far I have concentrated on empirically intended theories designed for experimental test, where a favourable test confers empirical adequacy. Not all theories are empirically intended. They are theories for which empirical adequacy isn't an intelligible goal. Think here of highly idealized normative theories in the manner of Bayesian treatments of belief revision or classical approaches to rational decision. The idealizations advanced by such theories—for example, that a rational agent will close his beliefs under consequence—are advanced in the certain knowledge of empirical discomportment with them. They don't describe how an actual reasoner actually reasons, but rather how an actual reasoner *should* reason or what he should try to approximate to in his reasoning. By far the hardest philosophical problem for model based normative theories is establishing the normative authority of its idealizations.[47] But what matters for us here is that it is widely thought that these theories are often successful and, when they are, they owe nothing of their success to an empirical adequacy they don't even seek. This raises a problem for my negative thesis, that is, for the claim that fictionality's detonation property wipes out all prospects of empirical adequacy. If a theory is a normative theory, how can it matter that fictionality has the detonation property? Detonation kills empirical adequacy, but empirical adequacy is not what good normative theories require or aspire to. Wouldn't this mean that there could be a place for fictionalism in normative models of belief revision, decision and the like?

No. Unlike empirically adequate theories in which the property of empirically infinite falsity doesn't detonate, in normatively idealized theories the property of empirical falsity *does* detonate. But here the detonation is (thought to be) compensated for by the normative authority of those derived falsehoods. Like all theories, a normative theory lives by its results. Theories generate conclusions intended for acceptance. We might schematize their structural arrangement as $T \vdash N_i$, where as before T is the theory, $\vdash$ is a consequence relation and the N_i are, with a certain contextual flexibility, T's "theorems". We take it as given that T's descriptively false normative idealizations are essential to the derivation of its theorems. Like the original idealizations, the theorems of T play a twofold role. They are descriptions of the behaviour of ideally rational agents, and they are not necessarily achievable norms for actual agents, for beings like us. Any theorem of T which depends on an idealized norm will itself take on that same normative texture. (Oughts in, oughts out.)

Suppose now that we exercised the fictionality option. In any serious application of them, fictions are made up. In literary cases they are created by authors. In scientific cases they are created by theorists, by modellers. There are few absurdities that won't be embraced by some philosopher or other. But the idea that the ideals of rationality are both normatively authoritative for you and me and anyone else who treads this planet, and yet free for the modeller's stipulation, is an idea

[47] See here Gabbay and Woods [53], See also Woods [54], Chap. 8.

that scarcely bears thinking about. Eddington is famous for saying that theories are put up jobs.[48] Of course, this was a joke. Eddington knew better than most that what you put a physical theory up to has to be paid for at the empirical checkout counter. Equally, no matter their stipulations, all the classical approaches to the normative modelling of rationality readily acknowledge that the bill for the makings up have to be paid somewhere. At a minimum, the model's idealizations would have to be descriptively adequate for ideally rational agency and thereby—it was supposed—normatively authoritative as a matter of objective fact for real-life reasoning or some plausible approximation of it. The fictionalization of the theory's theorems wipes out all prospect of meeting these objectives. One cannot make it descriptively accurate of ideal agency that belief is closed under consequence by putting on one's modelling hat and simply saying that it is. It is one thing to say that *in the model* ideal rationality is such-and-such. It is another thing entirely to say that the model is an accurate descriptor of perfect rationality. If fictionalization makes descriptive adequacy in idealized as regards real-life rationality an unachievable goal, so likewise it forecloses upon normative legitimacy in actual contexts. So we have learned another lesson.

> *Lesson* 5: If T is a normatively idealized theory, fictions undo its claims to normative authority.

14 Abductive Fictionalism

Disciplined and reflective thinking about science doesn't by any means always take the form of theories, whether syntactically or semantically construed. A shorter way of saying this is that not all exercises in scientific reasoning take the form, once completed, of structures such as $T \vdash O_i$. Most scientists are seized of the provisionality of even their most empirically well-favoured theories. Most scientists know that, whatever other properties it possesses $\vdash$ is not a relation of monotonic consequence. Our best theories to date lie exposed to the potential for damage occasioned by new information consistent with the original premisses but which, when added to them, snaps the $\vdash$-relation. On the other hand, some scientific reasoning is provisional in a much deeper way. It is reasoning of a kind that generates hypotheses for subsequent empirical test. Such reasoning has a broadly abductive character, discussed briefly by Aristotle under the name *apagogē*, translated as *abduction*. But it is to Peirce that we owe the modern invocation of it. In the best known of his scattered remarks on the subject, Peirce writes of abduction as follows:

[48] I owe the attribution to Quine in [55].

> 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.[49]

It is easy to see the Peirce's schema falls well short of a robust definition of abduction. For one thing, the schema embeds notions whose meanings, although intuitively familiar, are not precisely clear—"surprise", "matter of course", "reason to suspect". Even so, the schema gives unmistakable instruction about some of abduction's defining features, instruction which is reinforced in further passages of Peirce's work.[50] Peirce thinks that abduction is a form of guessing, and that a successful abduction provides no grounds for believing the abduced proposition to be true.[51] Rather than believing them, the proper thing to do with abduced hypotheses is to send them off to experimental trial.[52] Also important is that the connection between the abductive hypothesis and the observed fact is formulated subjunctively.[53] Similarly, the inference drawn from this subjunctive conditional is not that the abduced hypothesis is true but only that there is reason to suspect that it might be, and might be in a way that makes it a plausible candidate for empirical testing.[54]

Abduction is guessing. The most interesting epistemological fact about guessing is how good we are at it. It is the same way with abduction. We are good at it too. Perhaps the second most interesting epistemological fact about abductive guessing is how little is known of what enables us to be good at it. This is especially so when it comes to sorting out the conditions under which we shrink indefinitely large spaces of possible hypotheses to the one (or the few) that make the abductive cut. When, in his quest for a unified treatment of the laws of black body radiation, Planck thought up the quantum hypothesis, it was a proposition for which there wasn't a shred of antecedent evidence and none at all adduced by its presence as antecedent in the subjunctive conditional on which its provisional conjecture was based. Planck thought that the very idea of the quantum was bereft of physical meaning.

It is no condition on abductive adequacy that abduced hypotheses turn out well at experimental trial. There are more things whose truth was a reasonable thing to conjecture than actually turn out to be true. When Le Verrier conjectured the planet Vulcan he did so on the strength of the entirely defensible subjunctive conditional that if there were a heretofore undiscovered planet in that part of the heavens, Mercury's orbital perturbations would indeed be "a matter of course".

[49] Peirce [56], 5.189.

[50] For a recent analysis of Peircean abduction, see Woods [57]. This is a refinement and correction of an earlier treatment in Gabbay and Woods [58]. Also important are Aliseda [59], and Magnani [60]. An earlier treatment is Lipton's [61].

[51] Peirce [62].

[52] *Collected Papers*, 5.99; 6.49-6, 473; 7.202–219.

[53] *Collected Papers,* 5.189.

[54] *Collected Papers*, 5.189.

That was a sensible abduction at the time, notwithstanding that in the end it didn't pan out experimentally.

In some sense, the quantum hypothesis was down to Planck. Planck was the one who thought it up. Planck was the one who selected it for provisional engagement in a suitably adjusted physics. Some philosophers might see in these involvements a case for fictionalism. For aren't the sentences of fiction also down to their authors? Aren't the sentences of fiction the product of the author's thinking up, and of his own selection? When Conan Doyle thought up *The Hound of the Baskervilles* wasn't he imagining how things might have gone on the moors of western England in those years?

We have seen this point before. It is true that those and other similarities exist. But, again, what improvement in our understanding of physics would be achieved by calling the quantum hypothesis a fiction? It is well understood that if H is a working hypothesis in a theory, it is there on sufferance. It is on sufferance until such time that it earns its keep or is experimentally discredited. When the quantum hypothesis eventually paid off, it ceased being a hypothesis. Fictions aren't like this. Stories are not set-ups for subsequent experimental trial. That Holmes lived at 221B Baker Street is a fiction whose experimental test is entirely unmotivated and wholly untouched by a negative result if one were actually performed. Hypotheses are abduced. This is not the way in which fictions arise. How Doyle contrived Holmes' residency may not be entirely clear in all its details, but finding for "Holmes lived in Baker Street" a place as antecedent in a true subjunctive conditional of requisite abductive force is clearly not how it was done. This gives us a further lesson to draw:

Lesson 6: Abduced hypotheses aren't fictions.

15 Explanationist Fictionalism[55]

Some people think that Holmes was a psychopath. That I think is rather harsh, but let it pass. Suppose that Holmes *was* a psychopath and that this is a feature to which Doyle paid some attention. Suppose that some further Holmes stories have recently been discovered. In them Holmes' dark side is given careful and detailed scrutiny. It is possible, is it not, that in the story Doyle gives a plausible-seeming diagnostic account of his creation's affliction, an account in which hypotheses H_1 and H_2 play a role? (Doyle was himself a medical doctor.) Might not a reader of the stories seize upon these hypotheses and subject them to a scrutiny which his laboratory at King's College Hospital makes available to him? Couldn't such investigations turn out well for H_1 and H_2? Of course. Why then couldn't we say

[55] For explanationist fictionalism, see the three papers of Part III of Suarez [48] and the references therein: Elgin [63], Bokulich [64], and Morrison [65]. See also Bukulich [30].

that those stories served as hypotheses generators, hypotheses which, as it turned out, are true of the world?[56]

Yes, they did, but with a difference. Stories are subject to an anti-closed world presumption. Except where the author provides otherwise, the world of a story is the *world.* Stories would lack readers were this not so. Holmes lived in London. If the London in which Holmes resides bore no resemblance, except author-declared ones, to London, Holmes would be swallowed up in a swamp of indeterminacy, losing thereby any conceivable interest for even the most compliant reader. Holmes was a man. If the man that Holmes was bore no resemblance, except author-declared ones, to how men actually are, this stifling indeterminacy recurs. We are put, so to speak, in a state of massive amnesia with respect to Holmes. Who could possibly care about a man who neither has nor lacks an oesophagus, and most of humanity's other parts? What use to us, thanks to the author's failure to pronounce on it, is a man who neither had nor lacked a mother or who lives at no specific distance from Berkeley Square? Indeed, I daresay that there are some people whose knowledge of Late-Victorian and Edwardian London derives entirely from the Holmes stories of that period.[57]

It is true that novels can give us knowledge of the world. But a certain caution is now called for. Doyle's stories can't make it true *in the world* that *Holmes*' dark side is explained by the hypotheses in question. The diagnosis that is accurate for Holmes in the story is accurate for us in the world. Even so, that H_1 and H_2 work for Holmes is not sufficient reason to think that they work of us. If it turns out that they do work for us, it will not have been because they did the same for Holmes in the story. Doyle's stories make those hypotheses true of Holmes, not of us. Still, those hypotheses are available to us if we want them, just by reading the story. No one in his right mind would pay those hypotheses the slightest mind if the anti-closed world assumption for fiction weren't true. In the stories, the hypotheses turned out to be explanatory for Holmes under the conditions of life and circumstance his world placed him in. But auctorially contrived exceptions aside, Holmes' world is our world as it was then. It would be entirely surprising if what turns out to have been the case in Holmes' world weren't in most instances what turned out to be the case here. More particularly, any true generalization about men and about the London of the day will be instantiated by Holmes in default of Dole's provisions to the contrary.

Stories of this sort have a sort of explanatory value. What explains Holmes' darkness in the story explains our darkness in the here and now. Stories can be instructive in these ways. Scientific theories frequently exhibit this same virtue. Their goodness lies in the clarity that their explanations effect. It is sometimes supposed that such resemblances are sufficient cause to ascribe the fictional

[56] A real-world example: Freud made a psychoanalytic investigation of the character of *Gradiva*: *A Pompean Fantasy,* a novella by Wilhelm Jensen.

[57] Next to Dickens, no important writer reveals London's social complexities and physical textures better than Doyle.

character of explanatorily instructive stories to scientific theories possessing this same explanatory character. Upon reflection, however, there is nothing to be said for the idea. The theories that have caught the eye of fictionalists owe their explanatory force to the ineliminable presence of falsehoods. It is this combination of falsity and indispensability that attracts the attribution of fictionality. But the explanatory value of the diagnosis that pivots on H_1 and H_2 owes nothing whatever to any falsehood, notwithstanding their occurrences in Doyle's stories and their diagnostic success there. It is true in the story that they work for Holmes and true in the world that they work for us. The sentence "H_1 and H_2 explain such and so symptoms" is true in the story and true in the world. But "H_1 and H_2 explain Holmes' symptoms" is true in the story and false in reality. Its falsity in reality is occasioned by its reference to Holmes. It is not occasioned by the falsity of the diagnosis. The anti-closed world assumption being what it is, what might be true is that the stories played a causally stimulating role in getting our real world scientist thinking seriously enough about H_1 and H_2 to run his own laboratory tests. But there is nothing in the procedural manuals of experimental science that requires or leaves room for the recording of these causal provocations.[58] So, then, another lesson:

Lesson 7: If T is an explanatory theory, fictions have no constructive role there.

16 Suppressing Detonation

In the matter of highly idealized but empirically successful science, the Q that I claim to have found is that the property of fictionality distribute to dependent sentences, that fictionality detonates in semantic integration contexts. If this is so, Q is indeed a deal-breaker for the fictionalization of empirically adequate model based science. For if the property of empirical forlorn falsehood detonated in such theories, the observational consequences on which empirical adequacy depends would themselves be forlornly false, hence as far from truth as it is empirically possible to be. If the detonation property for fiction is indeed the Q that I seek, fictionalism is finished for empirically adequate idealized theories. It is stunning setback, both principled and cheaply attained. Perhaps this will strike some readers as too quick (and too easy) by half.

Certainly there are theorists of fiction who are not all disposed to accept the detonation property. There are different sources of this disinclination. One is the attachment shown by Meinongeans and some others to the idea that fictional objects are inherently and widely incomplete and, accordingly, that most sentences about them are without truth value (that is, are neither true nor false in the story). A

[58] Recall August Kekulé's vision of the chemical structure of benzene, occasioned in the hallucinatory grip of *delirium tremens.*

second source of this scepticism is a more general dissatisfaction with the anti-closed world assumption. The assumption assumes that most of what is true of a real entity of a given type will also be true of a fictional entity of that same type. If the assumption fails, the radical incompleteness thesis reasserts itself. A third reason to query the detonation claim arises from the truth-dyadicity thesis. Let's schematize "true in fiction" as **f** and "true in actuality" as **a**. There arise at once questions about the closure properties of sentences $\ulcorner \mathbf{f}(\Phi) \urcorner$, $\ulcorner \mathbf{f}(\chi) \urcorner$, $\mathbf{a}(\chi)$, and so on. A related matter is the extent to which **f** and **a** distribute through the truth functional connectives. A case in point: $\ulcorner \mathbf{f}(\Phi) \urcorner$ and $\ulcorner \mathbf{a}({\sim}\Phi) \urcorner$ might be all right separately. But, dialethicists apart, everyone agrees that we can't have $\ulcorner \mathbf{a}(\Phi \wedge {\sim}\Phi) \urcorner$, for any Φ, and some writers won't allow $\ulcorner \mathbf{f}(\Phi \wedge {\sim}\Phi) \urcorner$ either.[59] These uncertainties stimulate a readiness to crimp closure conditions for **f**-sentences and, even more aggressively for admixtures of **f**-sentences and **a**-sentences. The impact of these foreclosures range from a much reduced capacity for semantic integration to outright exclusion of it.

Perhaps these reservations are a seemly caution. Certainly there is no room here for dogmatism. But it would be wrong to think that the elimination of semantic integration contexts for fiction gives to philosophers of science of fictionalist bent cause to rejoice. It is quite true that if fictional sentences don't semantically integrate with real-world sentences, or do so only in some sternly crimped way, fictionality won't detonate in semantic integration contexts or will do so only to a sternly crimped degree. One way to block detonation is to block semantic integration. It is also true that semantic integration is essential for empirically adequate theoretical science. But semantic integration is not *sufficient* for a property's detonation. Whether a property detonates in a semantic integration context depends on which property it is. As we have seen, fictionality does and infinitely remote falsity doesn't. Whatever else we might say of it, a highly idealized but empirically adequate theory T is a semantic integration context. It is a semantic integration context in which the property of infinitely false idealization doesn't detonate. If semantic integration weren't an available context for fiction, it couldn't be true that the fictionality property detonates in semantic integration contexts. Whereupon, we have it that if infinitely false idealizations were fictions, T could not be a semantic integration context. But if that were so, T would lack the observational consequences on which its empirical adequacy depends.

On the other hand if fiction admitted of semantic integration under only tightly restricted conditions, there would be limited occasion for detonation of the fictionality property. Even this is too much for fictionalism to bear. If the idealizations of $T \vdash O_i$ were fictions, then in any circumstances in which part of T were a semantic integration context, infinitely remote falsity would distribute to the dependent O_i, wrecking their contribution to empirical adequacy. As for T's further parts, the parts that aren't semantic integration contexts, O_i couldn't be derived save from T's antecedently available observational sentences. In that case,

[59] Perhaps the most rigorous opponent of fictionally true inconsistencies is Lewis, in [66].

T's empirically verified observational consequences could reflect no dependency on the idealizations of T. Accordingly,

> *Lesson* 8: There is no relief for fictionalism in crimping T's semantic integration status.

17 A Brief Concluding Word

From early on I have recommended that the fictionalism question for science be handled without a large involvement in the logic of fiction. It has been satisfying to see that this is an achievable objective. Part of my disinclination to rely over-much on literary semantics is that the logic of fiction is surprisingly difficult, and the state in which we presently find it is riven by fundamental and dug-in disagreements. The contemporary record contains work of considerable ingenuity and sophistication. But the advice to avoid unnecessary work proved to be sensible and happily fulfillable.

It is no part of what I have been proposing here, still less of what I believe, that the literary semantics project be abandoned, that we capitulate to its difficulty and its captiousness. Fiction remains a standing, if as yet unresolved, challenge to a philosophically adequate semantics for natural language. To abandon it now would be intellectual dereliction. Even so, the fictionalist question for science requires neither its settlement nor its engagement.

It is possible, even so that actually contrary to what I have been saying here my Q-strategy and my pre-emption strategy haven't worked. It is possible that there is no Q that knocks scientific fictionalism out of the box. It is possible that there is such a Q but that it can be properly excavated only in a full-bore logic of fiction. Or it may be that fictionalism is true and that a true theory of fiction is somewhere to be had. But again, I ask, why would one look for it in the precincts of literary semantics? If a logic of fiction were necessary for implementing the requirements of the magnetic pull thesis for fictionalism in science, why wouldn't it serve us better to build one from the ground up?

Acknowledgments For criticisms and suggestions I thank a friendly audience in Sestri, including Atocha Aliseda, Daniel Arista, Howard Callaway, Michel Ghins, Michael Hoffman, Michael Lissak, Lorenzo Magnani, Woosuk Park, and Paul Thagard. I regret that I lack names for the other helpful intervenors. An earlier version of parts of this essay was delivered in Paris at the kind invitation of l'Institut Jean Nicod in December 2011. For challenging comments I am grateful to Jérôme Pelletier and Anouk Barberousse, and others whose names now escape me. Apologies. For conversation or correspondence before or after the conference, I also thank Nancy Cartwright, Nicholas Griffin, Dale Jacquette, Dominic McIvor Lopes, Shahid Rahman, and especially Alirio Rosales. For technical support I am deeply indebted—indeed tethered—to Carol Woods.

Appendix

Present-day theories of literary fictions reflect sharply different ways of cutting the cake. Here are two of them, made possible by subscription to or rejection of the following pair of assumptions:

> *Parmenides' Law*. Quantification and reference are existentially loaded. There is nothing that doesn't exist. It is not possible to refer to what isn't.
>
> *The Non-existence Postulate*. The purported objects and events of fiction do not exist. No object is a fictional object. No event is a fictional event.

It is doubtful that any philosophical claim could divide considered judgement more deeply than these two. Certainly they are, in their intractability, no improvement on the divisiveness occasioned by realism-antirealism wrangles in science or anywhere else. Why, then, for our philosophical anxieties about science, would we seek succor in the realist-antirealist war zones occasioned by the Law and the Postulate? Desperate times call for desperate measures, but isn't this going too far?

> A further point on which literary semanticists are divided is
>
> *Frege's Dismissal*. Since literature doesn't matter for science, a literary semantics would be of only marginal interest.[60]

In a rough and ready way, the first two of these clashing standpoints motivate the (incomplete) sample below, with the third somewhat orthogonal to them as in a second grouping, also just a sample.[61]

[60] "On *Sinn* and *Bedeutung*" translated by Max Black, in Michael Beaney editor *The Frege Reader*, pages 151-171, Oxford: Blackwell, 1997; p. 157. In "On denoting", Russell too gives fictional sentences the brush-off. They are, he says, sometimes true in a "secondary sense", without pausing to say what this sense might be. Strawson displays a similar casualness. In [67], he allows that sentences in *Pickwick Papers* are about Mr. Pickwick only in some (wholly unexplained) sense of "about".

[61] An excellent survey is Howell's "Literary fictions, real and unreal", in *Fictions and Models*, pages 27-107.

List One

Pro the law and the postulate	*Contra the law and the postulate*
Sayso semantics.[62]	Meinongean theories[63]
Pretense theories[64]	Existence-neutral logics[65]
Frege-Russell theories[66]	Artifactual theories[67]
Free logics[68]	Fictional worlds theories[69]
Theories of substitutional quantification[70]	

List Two

Pro Frege's dismissal	*Con Frege's dismissal*
Free logics	Sayso semantics
Frege-Russell theories	Meinongean theories
Strawson's "On Referring"	Pretense theories
	Artifactual theories
	Fictional worlds theories

There is in these multiplicities fair warning. As we have it now, the state of play in the logics of literary fictions give uncertain guidance to the realist-antirealist debate in science, or elsewhere. It is a conflicted matter in the philosophy of science. It is a conflicted matter in the philosophy of literature. So where in the philosophy of literature is the payoff for the philosophy of science?

[62] *The Logic of Fiction*, and "Fictions and their Logics".

[63] Richard Routley, *Exploring Meinong's Jungle and Beyond*, Canberra: Research School of Social Sciences, Australian National University, 1980, Parsons, *Nonexistent Objects,* 1980, and Jacquette, *A Meinongean Logic,* 1996, and Nicholas Griffin, "Through the Woods to Meinong's jungle", in Kent A peacock and Andrew D. Irvine, editors, *Mistakes of Reason: Essays in Honour of John Woods,* pages 15-32, Toronto: University of Toronto Press, 2005.

[64] Walton, *Mimesis as Make-Believe,* and David Lewis, "Truth in fiction".

[65] Routley [68] and Woods, *The Logic of Fiction*.

[66] Frege, "On *Sinn* and *Bedeutung*", Russell, "On denoting" and *An Introduction to Mathematical Philosophy.*London: Allen and Unwin, 1967. First published in 1919.

[67] Thomasson [69], and "Fiction, existence and indeterminacy", in Woods, *Fictions and Models*, and Juan Redmond, *Logique Dynamique de la Fiction: Pour un Approche Dialogique,* London: College Press, 2012.

[68] Lambert [70–72]. van Fraassen [73, 74], Burge [75], and Sainsbury [76].

[69] Woltersdorf [77], and Pavel [78].

[70] *The Logic of Fiction.*

References

1. Woods, J.: Fictionality and the logic of relations. South. J. Philos **7**, 51–64 (1969)
2. Woods, J.: The Logic of Fiction: Philosophical Soundings of Deviant Logic. Mouton, The Hague (1974). Reissued with a Foreword by Griffin, N., College Publications, London (2009)
3. Kripke, S.: Reference and Existence. Oxford University, Oxford (1973). Unpublished manuscript of the John Locke Lectures
4. Walton, K.: On fearing fictions. J. Philos. **75**, 5–29 (1978)
5. Howell. R.: Fictional objects and how they are and aren't. In: Woods, J., Pavel, T. (eds.) A Special Issue of Poetics, vol. 8, pp. 50–75 (1979)
6. Eaton, T.: Literary Semantics. The Melrose Press, Ely (2010)
7. Woods, J.: (ed.) Fictions and Models: New Essays, Foreword by Cartwright, N. volume one of Basic Philosophical Concepts, under the general editorship of Burkhardt, H., Philosophia, Munich (2010)
8. Suarez, M. (ed.): Fictions in Science: Philosophical Essays on Modelling and Idealization. Routledge, London (2009)
9. Bonevac, D.: Fictionalism. In: Irvine, A. D., (ed.) Philosophy of Mathematics, pp. 345–393, Handbook of the Philosophy of Science, Gabbay, D.M., Thagard, P., Woods, J. (eds.). North-Holland, Amsterdam (2009)
10. Magnani, L.: Scientific models are not fictions. Model-based science as epistemic warfare. In Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science. Western and Eastern Studies, pp. 1–38. Springer, Heidelberg (2012)
11. Woods, J., Isenberg, J.: Psychologizing the semantics of fiction. Methodos, online http://www.springer.com. (2010)
12. Cohen, M.: On the logic of fictions. J. Philos **20**, 477–488 (1923)
13. Walton, K.: Mimesis as Make-Believe. Harvard University Press, Cambridge (1990)
14. Parsons, T.: Nonexistent Objects. Yale University Press, New Haven (1980)
15. Jacquette, D.: Meinongean Logic: The Semantics of Existence and Nonexistence. De Gruyter, Berlin (1996)
16. Cartwright, N.: How the Laws of Physics Lie. Oxford University Press, Oxford (1980)
17. Bernecker, S., Pritchard, D. (eds.): Bayesian epistemology. The Routledge Companion to Epistemology, pp. 609–620. Routledge, London (2011)
18. Kant, I.: Logic. Bobbs-Merrill, Indianapolis (1974). First published in (1800)
19. Russell, B.: The Principles of Mathematics, 2nd edn. Allen and Unwin, London (1937), originally published in 1903
20. Fine, A.: Science fictions: comment on Godfrey-Smith. Philos. Stud. **143**, 117–125 (2009)
21. Woods, J.: Fictions and their logics. In: Jacquette, D. (ed.) Philosophy of Logic, pp. 1061–1126, a volume of the Handbook of the Philosophy of Science. Gabbay, D.M., Thagard, P., and Woods, J., (eds.). North-Holland, Amsterdam (2007)
22. Woods, J., Rosales, A.: Virtuous distortion: Idealization and abstraction in model- based science. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) Model-Based Reasoning in Science and Technology: Abduction, Logic and Computational Discovery, pp. 3–30. Springer, Berlin (2010)
23. Godfrey-Smith, P.: Models and fictions in science. Philos. Stud. **143**, 101–116 (2009)
24. Batterman, R.: Critical phenomena and breaking drips: infinite idealizations in physics. Stud. Hist. Philos. Sci. **36**, 225–244 (2005)
25. Batterman, R.: Idealization and modelling. Synthese **160**, 427–446 (2009)
26. Teller, P.: Fictions, fictionalisms and truth in science. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modelling and Idealization. Routledge, London (2009)
27. Giere, R.: Why scientific models should not be regarded as works of fiction. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modelling and Idealization. Routledge, London (2009)

28. Suárez, M.: Fictions, inference, and realism. In: Woods, J. (ed.) Fictions and Models: New Essays, Foreword by Cartwright, N., volume one of Basic Philosophical Concepts, under the general editorship of Burkhardt, H., Philosophia, Munich (2010)
29. Frigg, R.: Fictions and science. In: Woods, J. (ed.) Fictions and Models: New Essays, Foreword by Cartwright, N., volume one of Basic Philosophical Concepts, under the general editorship of Burkhardt, H. Philosophia, Munich (2010)
30. Bokulich, A.: How scientific models explain. Synthese **180**, 33045, 42 (2011)
31. Vaihinger, H.: The Philosophy of 'As If'. Odgen, C.K. translator, Routledge and Kegan Paul, London (1924). Originally published in German in (1911)
32. Quine, W.V.: Theories and Things. Harvard University Press, Cambridge (1981)
33. Woods, J., Rosales, A.: Unifying the fictional?. In: Fictions and Models, pp. 345–388. Volume one of Basic Philosophical Concepts, under the general editorship of Burkhardt, H., Philosophia, Munich (2010b)
34. Suppes, P.: A comparison of the meaning and uses of models in mathematics and the empirical sciences. Synthese **12**, 289–301 (1960)
35. Suppes, P.: Models of data. In: Nagel, E., Suppes, P., Tarski, A. (eds.) Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Conference, pp. 252–261. Stanford University Press, Stanford (1962)
36. McMullin, E.: Galilean idealization. Stud. Hist. Philos. Sci. **16**, 247–273 (1985)
37. Wigner, E.: The unreasonable effectiveness of mathematics in the natural sciences. In: Wigner, Symmetrics and Reflections: Scientific Essays, pp. 222–237. Indiana University Press, Bloomington (1967). Originally a Richard Courant lecture in mathematical sciences, New York University, 11 May 1959
38. Cartwright, N.: Foreword to Woods, J. (ed,) Fictions and Models, pp. 10-11. Volume one of Basic Philosophical Concepts, under the general editorship of Burkhardt, H. Philosophia, Munich (2010)
39. Fine, K.: The pure logic of ground. Rev. Symb. Log. **5**, 1–25 (2012)
40. Van Fraassen, B.:The Image of Science. Clarendon Press, Oxford (1980)
41. Rosen, G.: What is constructive empiricism? Philos. Stud. **94**, 143–178 (1994)
42. Russell, B.: Introduction to Mathematical Philosophy. Routledge and Kegan Paul, London (1919). Reprinted in 1993
43. Burgess, J.: Fixing Frege. Princeton University Press, Princeton (2012)
44. Bentham, J.: Bentham's theory of fictions. In: Ogden, C.K., (ed.) Routledge and Kegan Paul, London (1932)
45. Bentham, G.: Outline of a New System of Logic. Thoemmes, Bristol (1990). First published in 1827
46. Quine, W.V. : Five milestones of empiricism. In Quine: Theories of Things, pp. 67–72. Harvard University Press, Cambridge (1981)
47. Bentham, J.: De l'Ontologie et autres textes sur les fictions. Schofield, P., Cléro, J.P., and Laval, C., (eds.) Point Seuil, Paris. Appendix B, p. 164 (1997)
48. Ramsey, F.: Theories. In his Foundations of Mathematics and Other Logical Essays, Braithwaite, R.B., (ed.) With a preface by Moore, G.E. Routledge and Kegan Paul, London (1931)
49. Craig, W.: On axiomatizability within a system. J. Symb. Log. **17**, 30–32 (1953)
50. Craig, W.: Replacement of auxiliary expression*s*. Philos. Rev. **55**, 38–55 (1956)
51. Craig, W.: Bases for first-order theories and subtheories. J. Symb. Log. **15**, 97–142 (1960)
52. Craig, W.: Elimination theorems in logic: a brief history. Synthese **164**, 321–332 (2008)
53. Gabbay, D.M., Woods, J.: Normative models of rational agency: the disutility of some approaches. Log. J. IGPL **11**, 597–613 (2003)
54. Woods, J.: Paradox and Paraconsistency: Conflict Resolution in the Abstract Sciences. Cambridge University Press, Cambridge (2003)
55. Quine, W.V.: Homage to Carnap. Boston Studies in Philosophy of Science, vol. 8, pp. xxii–xxv. Reidel, Dordrecht (1971)
56. Peirce, C.S.: Collected Papers. Harvard University Press, Cambridge, (1931–1958)

57. Woods, J.: Cognitive economics and the logic of abduction. Rev. Symb. Log. **5**, 148–161 (2012)
58. Gabbay, D.M., Woods, J.: The Reach of Abduction: Insight and Trial. A Practical Logic of Cognitive Systems, vol. 2. North-Holland, Amsterdam (2005)
59. Aliseda, A.: Abductive reasoning: logical investigations into the processes of discovery and evaluation. Springer, Amsterdam (2006)
60. Magnani, L.: Abductive Cognition: The Epistemology and Eco-Cognitive Dimension of Hypothetical Reasoning. Springer, Heidelberg (2009)
61. Lipton, P.: Inference to the Best Explanation, 2nd edn. 2004. Routledge, London (1991)
62. Peirce, C.S.: Reasoning and the Logic of Things: The Cambridge Conference Lectures of 1898, p. 178. Harvard University Press, Cambridge (1992)
63. Elgin, C.Z.: Exemplification, idealization, and scientific understanding. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modelling and Idealization, pp. 79–90. Routledge, London (2009)
64. Bokulich, A.: Explanatory fictions. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modelling and Idealization, pp. 91–109. Routledge, London (2009)
65. Morrison, M.: Fictions, representations, and reality. In: Suarez, M., (ed.), Fictions in Science: Philosophical Essays on Modelling and Idealization, pp. 120–135. Routledge, London (2009)
66. Lewis, D.: Truth in fiction. Am. Philos. Q. **15**, 37–46 (1978)
67. Strawson, P.F.: On referring. Mind **59**, 320–344 (1950)
68. Routley, R.: Some things do not exist. Notre Dame J. Formal Log. **7**, 251–276 (1966)
69. Thomasson, A.: Fictions and Metaphysics. Cambridge University Press, Cambridge (1998)
70. Lambert, K.: Notes on E! III: a theory of descriptions. Philos. Stud. **13**, 51–59 (1963)
71. Lambert, K.: Existential import revisited. Notre Dame J. Formal Log. **4**, 288–292 (1963)
72. Lambert, K.: Notes on E! IV: A theory of Descriptions. Notre Dame J. Formal Log. **15**, 85–88 (1964)
73. van Fraassen, B.C.: The completeness of free logic. Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik **12**, 219–234 (1966a)
74. van Fraassen, B.C.: Singular terms, truth-value gaps, and free logic. J. Philos. **63**, 481–494 (1966b)
75. Burge, T.: Truth and singular terms. Noûs **8**, 309–325 (1974)
76. Sainsbury, M.: Fiction and Fictionalism. Routledge, London (2009)
77. Woltersdorf, N.: Works and Worlds of Art. Oxford University Press, New York (1980)
78. Pavel, T.: Fictional Worlds. Harvard University Press, Cambridge (1986)

Fear-Driven Inference: Mechanisms of Gut Overreaction

Paul Thagard and A. David Nussbaum

Abstract Model-based reasoning requires not only inferences about what is happening, but also evaluations of the desirability of what is happening. Emotions are a key part of such assessments, but sometimes they can lead people astray, as in motivated inference when people believe what fits with their desires. In contrast to motivated inference, fear-driven inference generates beliefs that people do not want to be true. Although paradoxical, this kind of inference is common in many domains, including romantic relationships, health, parenting, politics, and economics. This paper proposes that fear-driven inference results from gut overreactions, in which a feeling that something is wrong is erroneously taken as evidence that something really is wrong. We discuss psychological and neural mechanisms by which gut overreactions can lead to fear-driven inference, and show how a computer model of emotional coherence can explain both fear-driven and motivated inference.

1 Introduction

> Trifles light as air are to the jealous confirmations strong (Iago, in *Othello).*

In Shakespeare's play, Othello is led on the basis of flimsy evidence to conclude that his wife Desdemona is unfaithful to him. This belief is highly distressing to him, but he cannot help becoming increasingly convinced by a supposition that he does not want to be true. Othello's conclusion is an instance of *fear-driven*

P. Thagard (✉)
University of Waterloo, Waterloo, Canada
e-mail: pthagard@uwaterloo.ca

A. D. Nussbaum
University of Chicago, Chicago, USA
e-mail: a.nussbaum@chicagobooth.edu

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_3, © Springer-Verlag Berlin Heidelberg 2014

inference, in which people believe something, not just despite the fact they fear it to be true, but partly *because* they fear it to be true. This article identifies instances of fear-driven inference in several domains, and proposes psychological and neural mechanisms that explain how people can paradoxically be inclined to believe just what they least want to believe.

Researchers have identified many kinds of cognitive and emotional biases that distort human thinking, such as representativeness, anchoring, confirmation bias, and motivated inference [1–3]. Motivated inference occurs when people reach conclusions unduly driven by their personal goals rather than by the available evidence [4–6]. This kind of thinking might be called *desire-driven* inference, but it is more complex than mere wishful thinking in which people believe something just because they want it to be true. Motivated inference is more subtle in that desires lead to prejudiced selection and weighting of evidence.

Elster [7] has noticed a much less well-known kind of biased inference in which people reach conclusions that go against their desire. He calls it *countermotivated* inference, indicating that people make inferences contrary to their motivations. We propose the term *fear-driven inference* because it points to the kinds of psychological and neural mechanisms based on negative emotions that can lead people to reach conclusions that run contrary to their interests. In the nineteenth century, Mill [8, pp. 482–483] recognized this kind of inference in his *System of Logic* when he wrote:

> The most common case [of bias] is that in which we are biased by our wishes; but the liability is almost as great to the undue adoption of a conclusion which is disagreeable to us, as of one which is agreeable, if it be of a nature to bring into action any of the stronger passions. Persons of timid character are the more predisposed to believe any statement, the more it is calculated to alarm them. Indeed it is a psychological law, deducible from the most general laws of the mental constitution of man, that any strong passion renders us credulous as to the existence of objects suitable to excite it.

Such thinking was recognized even earlier in the fable "Le Loup et le Renard" by the seventeenth-century writer Fontaine [9], who wrote: "Chacun croit fort aisément ce qu'il craint et ce qu'il désire." Mele [10] used the term "twisted self-deception" for self-deception that involves the acquisition of an unwanted belief, another kind of fear-driven inference.

2 Emotions and Rationality

Are people rational or emotional? Much recent research in psychology, neuroscience, and economics challenges the dichotomous presupposition of this question, showing that emotional reactions to situations are often a key part of rationality. Discussions of feeling as information [11, 12], emotional intelligence [13], the affect heuristic [14], and the somatic marker hypothesis [15] all describe ways in which emotional reactions can efficiently summarize complex evaluations of situations and provide a guide to action.

However, the recognition of emotion as often a valuable contributor to rationality should not obscure the many occasions when emotions contribute to irrational behavior. Obvious examples include the irrational exuberance [16] of financial bubbles such as the dot.com boom of the 1990s and the housing debacle of the mid-2000s, as well as destructive emotions such as racial hatred and cravings for addictive drugs. In order to sort out the ways that emotions can contribute to rational and irrational thinking and behavior, we need to gain a deeper understanding of how emotions are an integral part of judgment and decision making.

This paper identifies a pattern of emotional irrationality that we call *gut overreaction,* in which an amplifying feedback loop between judgments and emotions can lead both to excessively positive assessments and to excessively negative ones. Such overreactions occur in many spheres of human life, unduly influencing financial decisions, personal relationships, and medical behavior. We will first illustrate the operation of gut overreactions in *fear-driven inference*. In common cases in personal relationships including both romantic and parental ones, people experience irrationally negative emotions. Then we will show how the same underlying neuropsychological mechanism of gut overreaction can lead to irrationally positive emotions of the sort that occur in financial bubbles and romantic infatuation. Finally, we discuss psychological and social techniques for avoiding and overcoming the irrational results of gut overreactions.

3 Fear-Driven Inference

Shakespeare's Othello becomes obsessively worried that his wife Desdemona is unfaithful to him, despite the flimsiness of evidence planted by the evil Iago. Such irrational jealousy is sufficiently common that it has been dubbed the *Othello syndrome* [17], also known as morbid jealousy. What is amazing about this pattern of thinking is that it is doubly irrational, going both against the available evidence and against the best interests of the reasoner. Not only does Othello have more evidence that his wife loves him than that she is cheating, he is made deeply miserable by the thought that she is cheating on him. Why would people go against both the evidence and their own interests?

The Othello syndrome may be rare, but a similar phenomenon occurs in many parents of adolescents. Parents naturally worry about what their teenagers are up to, and failures to call or return home when expected may prompt parents to intense anxiety about what might have happened to their children. Such anxiety is commonly recognized after the fact as excessive, if parents have adequate evidence from their children's previous behavior to infer a benign explanation for their current lapses. As in Othello's jealousy, overanxious parents engage in inference that goes against the available evidence that probably nothing bad has happened to their children and, also against the parent's own self-interest of being calm and confident. Such parental anxiety is both evidentially irrational and highly unpleasant.

Fear-driven inference arises in many other domains that are sufficiently important to people to generate anxiety. For example, people naturally care about their health, which can lead them to think they are more sick than they actually are. Hypochondriacs (and even ordinary people such as medical students whose training acquaints them with hundreds of obscure diseases) may infer from some minor symptom that they have a serious disease, without considering the full range of evidence and alternative hypotheses. People believe that they have a disease not just despite the fear that they have the disease, but because of the fear.

Other instances of fear-driven inference occur in thinking about careers, economics, politics, and religion. An academic who submits a paper for publication and gets no response to it for a long time may start to infer that the journal is just not interested in the paper, even though there are many other explanations for editorial delay. Investors may swing from irrational exuberance about stocks or other financial concerns to irrational despair that results from the fact that they fear financial disaster. Fear-driven inference is rampant in politics as seen in the popularity of conspiracy theories and other kinds of paranoia: people are sometimes inclined to believe the worst because it scares them, although motivated inference can also contribute to beliefs in conspiracies. Finally, belief in religion is often supported by motivated inferences concerning benign gods or blissful afterlives [18], but it can also be fear-driven when inspired by worries about vengeful deities and eternal punishment. In both motivated and fear-driven inference, feelings are *mis*information.

4 Gut Overreaction

Why are people prone to fear-driven inference? It is much easier to understand the psychological basis of *motivated inference,* in which people distort their judgments because of their underlying personal goals [5, 6]. Motivated inference is an emotional bias that undercuts rationality, and can be observed in many kinds of interpersonal and practical judgments. For example, people buying lottery tickets may understand that the expected value of winning is very low, but nevertheless be convinced that this is their lucky day. The underlying psychological mechanism of motivated inference may be a kind of emotional coherence in which our goals and values naturally but illegitimately influence what we come to believe [18, 19]. But emotional coherence cannot explain cases such as the Othello syndrome and parental overanxiety, where the distressing emotional results clearly go against the goals of the worriers. A different psychological mechanism must be at work.

We propose that the mechanism underlying fear-driven inference is gut overreaction, which involves an ongoing feedback loop between judgment and emotional response. Current emotion theories tend to divide into two camps, one that considers emotions to be akin to judgments [20], and the other that considers

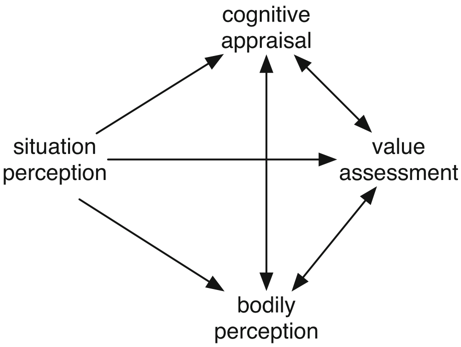


Fig. 1 Emotion as an integrated process of assessment of value deriving from both cognitive appraisal and bodily perception

emotions to be reactions to bodily states [21]. This division, however, can be reconciled by considering the brain as reacting to situations in ways that take into account both cognitive appraisal of situations and perception of bodily states, as in the EMOCON model of Thagard and Aubie [22, 23]. A simplified version of this model, omitting neural details, is shown in Fig. 1. From this perspective, jealous spouses and anxious parents are experiencing worry because of both their appraisal of their situations *and* their internal perception of bodily changes.

If emotions involve neural integration of both cognitive appraisal and physiological perception, then it becomes evident how gut overreaction can occur. Consider the feedback loop shown in Fig. 2a, intended to explain Othello's fear-driven inference that Desdemona is unfaithful to him. The suggestion due to Iago's misinformation may lead Othello to suspect that Desdemona is cheating, but this makes him feel bad which in turn makes him even more suspicious of her. That she is unfaithful causes him to feel bad about her, which in an ongoing loop makes him more suspicious of her. The general case is shown in Fig. 2b, which applies equally well to parental anxiety. Thinking that things are bad (with children or anything else that matters) causes you to feel bad, which in turn leads you to become more convinced that things are bad. The amplifying interactions shown in Fig. 2 are usually described as a positive feedback loop, but we avoid that term "positive" here because of confusion with positive emotions. In fear-driven inference, the amplifying feedback loop between inference and emotional reaction leads to negative emotions such as anxiety and anger. Let us now consider how gut overreaction can also produce excessively positive emotions.

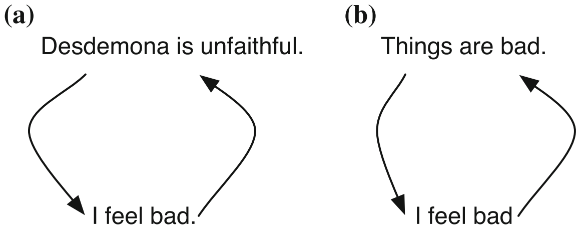


Fig. 2 Amplifying feedback loop producing negative emotions

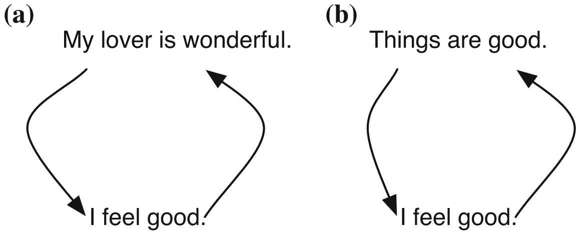


Fig. 3 Amplifying feedback loop producing positive emotions

5 Positive Overreactions

The early stages of romantic love are often attended by wildly enthusiastic emotional experiences such as obsessive feelings of joy and passion [24]. We conjecture that such infatuation is the result of the kind of amplifying feedback loop shown in Fig. 3a, in which the judgment that a romantic object is wonderful makes someone feel good, and the feeling itself is taken as support for the judgment that the loved one really is wonderful. The result can be an exaggeratedly positive attitude that may lead to disillusionment, or in a happier course of romantic development, to a more stable sort of companionate love that can develop after a year or so of infatuation.

Figure 3b shows the general pattern, which applies to many phenomena ranging from financial bubbles to religious experience. In an economic boom in stocks, housing, or commodities, prices keep going up and up. Cool heads advise that what goes up must come down, but they are ignored in what the economist Shiller [16] called *irrational exuberance*. This description was originally applied to the dot.com boom of the late 1990s, but fits equally well the housing and financial bubbles of the 2000s. Ideally, people making a decision whether to buy a stock, house, derivative, or commodity should do a duly diligent assessment of the probable costs and benefits of the purchase. But in a highly complex world such assessments are difficult to make, so people naturally fall back on their "gut reactions" that tell them how they feel about the purchase. When such emotions are based on a wealth of accumulated experience, the gut reaction can constitute a reasonable judgment. But the amplifying feedback loop shown in Fig. 3a shows how the emotional estimation of the purchase can fail to reflect reality, if people feel good about the opportunity because of their judgments, but their positive judgments are largely tied to their feeling good.

In both romantic infatuation and financial bubbles, irrational exuberance can be a group phenomenon, in which one person's exuberance feeds backs into another's, as shown in Fig. 4. If the people can perceive each other directly, then the interpersonal emotional feeback can involve mechanisms such as emotional contagion [25] or activation of mirror neurons [26]. Alternatively, social feedback can be indirect, as in stock market prices. Either way, amplifying social feedback increases the amplifying psychological feedback shown in Fig. 3.

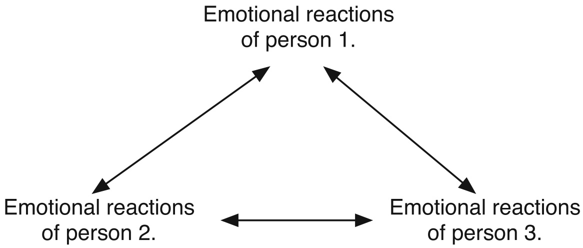

Emotional reactions of person 1.
Emotional reactions of person 2.
Emotional reactions of person 3.

Fig. 4 Social amplifying feedback loops produce spread of emotions

If we are correct, then gut overreaction is one of the psychological mechanisms underlying financial bubbles, as people feel better and better about feeling better and better. Of course, it is not the only relevant mechanism, as people's emotional reactions also derive from the perfectly reasonable recognition that prices have been going up, and from the motivated inference that prices will go up because they want them to go up. But the amplifying feedback loop between judgment and feeling can intensify and prolong the conviction that things can only get better.

Sadly, when things turn sour in the economy or personal relationships, people can swing from positive gut overreaction to negative overreaction, when the one amplifying feedback loop is supplanted by another. This transformation occurs when a blowup shifts romantic infatuation into disillusionment, and when a financial crash swings a bubble into an economic crisis. Akerlof and Schiller [27] discuss the importance of having a "confidence multiplier" operating in an economy, in which confidence breeds confidence and despair breeds despair. Gut overreaction may be one of the psychological mechanisms underlying this multiplier. Figure 5 illustrates the transition that can take place in people when events and new information cause a critical transition from motivated inference to fear-driven inference, producing a swing from irrational exuberance to excessive despair.

6 Computer Simulation of Fear-Driven Inference

In order to explore the effects of gut overreactions on inference, we have performed computer simulations of the effects of amplifying feedback loops on inferential dynamics. Consider the highly simplified version of Othello's case

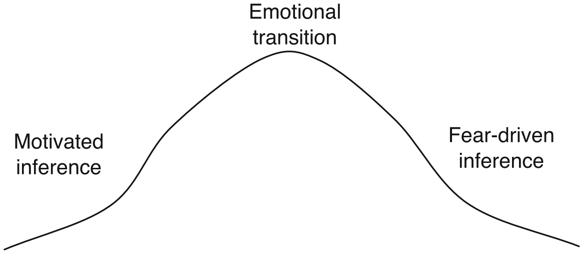


Fig. 5 Emotional transition resulting from shift away from motivated inference to fear-driven inference, turning financial or romantic bubbles into busts

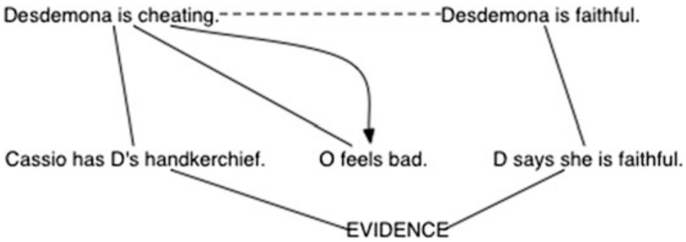


Fig. 6 Structure of neural network simulation of Othello's fear-driven inference. The *dotted line* indicates an inhibitory connection between two contradictory hypotheses. The straight *solid lines* are excitatory connections based on evidence relations. The *curved pointed line* indicates an emotional connection

shown in Fig. 6. In this simulation, there are two pieces of evidence: Desdemona's handkerchief has turned up in the possession of Cassio, yet Desdemona says she is faithful. That Desdemona is cheating with Cassio explains his having her handkerchief, and the contradictory hypothesis that she is faithful explains why she says she is faithful. Just based on this information, there is no reason to infer either that she is cheating that or she is faithful, and simulation using the neural network simulator HOTCO (hot coherence, [18]) yields equal activation for both hypotheses. However, there is a different result when HOTCO adds the node for feeling bad, which gets emotional activation from its association with the node that Desdemona is cheating. Then feeling bad irrationally becomes evidence that gets explained by the hypothesis that Desdemona is cheating, which becomes a kind of self-supporting hypothesis. In this way, the HOTCO simulation produces fear-driven inference by a kind of gut overreaction. In an alternative simulation, an "O feels good" node could support the motivated inference that Desdemona is faithful. Whether the "O feels good node" or the "O feels bad" node becomes activated can depend on many factors including social circumstances such as conversations and personality components such as neuroticism and low self-esteem.

7 Neural Mechanism

From the perspective of conventional theories of rationality based on probability and utility theory, fear-driven inference is bizarre. Theories of rationality assume a firewall between probabilities and utilities, which are calculated independently of each other and only brought together in calculations of expected utility through the classic equation that multiplies them. The brain, however, does not appear to separate probabilities and utilities nearly so rigorously. Evidence for interactions comes from psychological experiments that show that people use emotions to estimate probabilities [28] and that motivations affect belief [5]. What are the neural mechanisms that produce mingling of probabilities and utilities and can yield fear-driven inference?

Little research has been done on the neural correlates of belief, but Sam Harris and his colleagues have some interesting preliminary results [29, 30]. They found that the neural correlates of belief included brain areas associated with emotional processing, such as the ventromedial prefrontal cortex (for belief) and the anterior insula (for disbelief). Hence it is not surprising that the brain confuses emotional arousal with believability. In the case of scary hypotheses such as spousal infidelity, offspring mishaps, medical threats, or economic dangers, the arousal generated by fear may be confused with arousal generated by conviction based on evidence. Important domains such as family relations, health, and economic well-being naturally yield high activations in brain areas relevant to processing emotional information. Such activations can lead to fear-driven inferences when arousal is misinterpreted as probability rather than as disutility.

It is actually a strength, not a weakness, of the brain that it integrates cognition with emotion. Emotions provide focus on what is important to an organism, serving such important roles as ensuring that inferences will be made about goal-relevant information rather than about trivialities and providing an immediate connection between belief and action. Modern mathematical theories of probability and utility only arose in the seventeenth and eighteenth centuries [31], so it is not surprising that human thinking still often relies on cruder non-differentiated processes of belief assessment.

8 Controlling Gut Overreactions

If gut overreaction is a contributing factor to irrationality, what can be done about it? If our analysis is correct, there are many psychological and social ways to reduce the excessive effects of negative and positive emotions. Because bodily perception is part of the genesis of emotional reactions according to the EMOCON model, physiological interventions such as meditation are appropriate. Also potentially useful are drugs that alter levels of neurotransmitters, such as anti-anxiety medications and anti-depressants. At the more cognitive level, people can ask themselves: am I feeling good (or bad) about X because it really is good? The neural processes involved in emotional reactions to a situation are mostly inaccessible to conscious control, but techniques such as cognitive therapy can be used to examine the basis of the appraisals that are one of the factors that go into emotional reactions. Ideally, in keeping with the finding about depression that the best treatment involves both medication and therapy, attempts to modulate gut overreactions should operate both physiologically and cognitively. One useful tool for identifying the emotional background to inferences is the technique of cognitive-affecting mapping which displays the emotional values and connections of key concepts ([32], ch. 17).

Also important are social processes shown in Fig. 4. Group members can have amplifying feedback influences on each other, but other people who are less prone to overreaction or less involved can have dampening effects on an individual's own tendency to become excessively exuberant or despondent about a situation.

9 Conclusion

We have conjectured that gut overreactions produced by an amplifying feedback loop between judgments and emotions can be an important factor in many kinds of irrationality operating in spheres that range from personal relationships to economic dynamics. Wishful thinking, understood at a deeper psychological level as motivated inference, has an important counterpoint in fearful thinking, which we have analyzed as fear-driven inference deriving from gut overreactions. Such overreactions produce feelings as misinformation. Much research remains to be done to provide evidential evaluation concerning the neural feedback processes we have hypothesized and concerning the psychological effects of these processes.

Acknowledgments Thagard's research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Gilovich, T.: How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life. Free Press, New York (1991)
2. Gilovich, T., Griffin, D., Kahneman, D. (eds.): Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, Cambridge (2002)
3. Kahneman, D., Tversky, A. (eds.): Choices, Values, and Frames. Cambridge University Press, Cambridge (2000)
4. Bastardi, A., Uhlmann, E.L., Ross, L.: Wishful thinking: belief, desire, and the motivated evaluation of scientific evidence. Psychol. Sci. **22**, 731–732 (2011)
5. Kunda, Z.: The case for motivated reasoning. Psychol. Bull. **108**, 480–498 (1990)
6. Kunda, Z.: Social Cognition: Making Sense of People. MIT Press, Cambridge (1999)
7. Elster, J.: Explaining Social Behavior. Cambridge University Press, Cambridge (2007)
8. Mill, J.S.: A System of Logic, 8th edn. Longman, London (1970)
9. de la Fontaine, J.: Le loup et le renard. http://www.lafontaine.net/lesFables/afficheFable.php?id=214 (2012)
10. Mele, A.R.: Self-Deception Unmasked. Princeton University Press, Princeton (2001)
11. Schwartz, N., Clore, G.L.: Mood as information: 20 years later. Psychol. Inq. **14**, 296–303 (2003)
12. Schwarz, N.: Feelings as information: informational and motivational functions of affective states. In: Higgins, E.T., Sorrentino, R. (eds.) Handbook of Motivation and Cognition: Foundations of Social Behavior, pp. 527–561. Guilford Press, New York (1990)
13. Salovey, P., Detweiler-Bedell, B.T., Detweiler-Bedell, J.B., Mayer, J.D.: Emotional intelligence. In: Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (eds.): Handbook of Emotions, 3rd edn, pp. 533–547. Guilford Press, New York (2008)

14. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: The affect heuristic. In: Gilovich, T., Griffin, D., Kahneman, D. (eds.) Heuristics and Biases: The Psychology of Intuitive Judgement, pp. 397–420. Cambridge University Press, Cambridge (2002)
15. Damasio, A.R.: Descartes' Error: Emotion, Reason, and the Human Brain. G. P. Putnam's Sons, New York (1994)
16. Schiller, R.: Irrational Exuberance, 2nd edn. Princeton University Press, Princeton (2005)
17. Todd, J., Dewhurst, K.: The Othello syndrome: a study in the psychopathology of sexual jealousy. J. Nerv. Ment. Dis. **122**(4), 367–374 (1955)
18. Thagard, P.: Hot Thought: Mechanisms and Applications of Emotional Cognition. MIT Press, Cambridge (2006)
19. Thagard, P.: Why wasn't O. J. convicted? Emotional coherence in legal inference. Cogn. Emot. **17**, 361–383 (2003)
20. Scherer, K.R., Schorr, A., Johnstone, T.: Appraisal Processes in Emotion. Oxford University Press, New York (2001)
21. Prinz, J.: Gut Reactions: A Perceptual Theory of Emotion. Oxford University Press, Oxford (2004)
22. Thagard, P., Aubie, B.: Emotional consciousness: a neural model of how cognitive appraisal and somatic perception interact to produce qualitative experience. Conscious. Cogn. **17**, 811–834 (2008)
23. Thagard, P.: The Brain and the Meaning of Life. Princeton University Press, Princeton (2010)
24. Fisher, H.: Why We Love: The Nature and Chemistry of Romantic Love. Henry Holt, New York (2004)
25. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Emotional Contagion. Cambridge University Press, Cambridge (1994)
26. Iacoboni, M.: Mirroring People: The New Science of How We Connect with Others. Farrar, Straus and Giroux, New York (2008)
27. Akerlof, G.A., Shiller, R.J.: Animal Spirits: How Human Psychology Drives the Economy, and why it Matters for Global Capitalism. Princeton University Press, Princeton (2009)
28. Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N.: Risk as feelings. Psychol. Bull. **127**, 267–286 (2001)
29. Harris, S., Kaplan, J.T., Curiel, A., Bookheimer, S.Y., Iacoboni, M., Cohen, M.S.: The neural correlates of religious and nonreligious belief. PLoS One **4**(10), e7272 (2009). doi: 7210.1371/journal.pone.0007272
30. Harris, S., Sheth, S.A., Cohen, M.S.: Functional neuroimaging of belief, disbelief, and uncertainty. Ann. Neurol. **63**, 141–147 (2008)
31. Hacking, I.: The Emergence of Probability. Cambridge University Press, Cambridge (1975)
32. Thagard, P.: The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change. MIT Press, Cambridge (2012)

Living in the Model: The Cognitive Ecology of Time—A Comparative Study

Chris Sinha

Abstract Time is at once familiar and mysterious, its status in the physical universe being uncertain and contested. Time seems to be fundamental to both biology and to the world of human experience. It seems certain that human beings in all cultures experience time, and have ways of linguistically referring to relations between events in time. It has been proposed by some cognitive scientists that there is a natural, transcultural conceptual domain of time. Cultural conceptions of time, however, vary considerably. I present anthropological linguistic data from a study that my colleagues and I conducted in an indigenous Amazonian community. Concepts of time are cultural and historical constructions, constituted by schematic time interval systems, and embodied in language and culture dependent symbolic cognitive artefacts. "Living in time", I contend, is to live in a model. Time is both artifactual model and cognitive niche, made possible by the wider biocultural niche of language.

1 Introduction: Time in Cosmos, Life and Mind

Time is familiar, but not something we usually think of as a friend. It is not just that it is part of our everyday lives. Time seems somehow to be above and beyond everyday life, giving it pattern and structure, intermeshing it with the lives of

My heartfelt thanks go to my collaborators in the field-based case study described and analyzed in this chapter, Wany Sampaio, Vera da Silva Sinha and Jörg Zinken; to the Amondawa community, for allowing us to share their language and culture; and to the Editor of this volume, Lorenzo Magnani, for his patience and support. I hope that the ideas outlined in this paper meet with the approval of these colleagues and partners in dialogue, while taking sole responsibility for all its defects.

C. Sinha (✉)
School of Languages and Literature, Lund University, Lund, Sweden
e-mail: chris.sinha@ling.lu.se

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_4, © Springer-Verlag Berlin Heidelberg 2014

others. The measure of time governs our every waking moment, time is precious, its passage is relentless, what remains of it is ever-diminishing. But time is also mysterious. Its nature, even its very existence, defy consensus in cosmology, physics and philosophy.

Isaac Newton [1] believed time, like space, to be absolute and infinite: "Absolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly and by another name is called duration. Relative, apparent, and common time is any sensible and external measure (precise or imprecise) of duration by means of motion; such a measure—for example, an hour, a day, a month, a year—is commonly used instead of true time". In this famous passage, Newton asserted the metaphysical reality of Time as an independent dimension of the universe, a position that commanded consent for centuries until challenged, almost simultaneously, in physics by Einstein's Special Theory of Relativity [2] and in philosophy by McTaggart [3]. Newton also availed himself of a metaphor, of the "flow" (or "passage") of the River of Time [4], whose status would later prove as troublesome for relativistic anthropological linguistics as it had for relativistic physics. Benjamin Lee Whorf formulated what he called "The Principle of Linguistic Relativity" [5] on the basis of his analysis of concepts of time and temporality in the Native American Hopi language. The Hopi speaker, he said, "has no general notion or intuition of *time* as a smooth flowing continuum in which everything in the universe proceeds at an equal rate, out of a future, through a present, into a past; or, in which, to reverse the picture, the observer is being carried in the stream of duration continuously away from a past and into a future" [6].

Einstein himself believed time to be an illusion, engendered by consciousness: in the relativistic space–time Block Universe, past and future are equally real, and an objective present does not (and cannot) exist. As Einstein's contemporary, the mathematician Minkowski put it, "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality" [7].

Not everyone is convinced by the denial of "passage" in the deterministic Block Universe. The reality of the Arrow of Time—the reality, that is, of irreversibility, whether in thermodynamics, in biological evolution or in the individual lifespan—has been invoked by many theorists in defence of the reality of time and the present moment, and of the felt asymmetry of past and future. The idea that the future is as determinate as the past is difficult to reconcile with a fundamental condition of the intelligibility of our social lives: that we are, as agents, accountable for our past acts, but cannot be held responsible for that which we have not (yet?) done.

Unfortunately, the relationship between determinism and indeterminism in physics is no more settled than the problem of time itself. It certainly lies way beyond the scope of this chapter to attempt any resolution of the problem of time in general. My aim is more modest: to explore the way our concepts of time and temporality have emerged in history and culture as fundamental constituents of human cognitive ecology; and to make strange the taken-for-grantedness of these concepts by inviting the reader to view them through the lens of the very different understanding of time in an indigenous Amazonian culture. At the end of this journey in time and space, after our brief encounter with Amazonian time, I will address again Whorf's question of the relation between language and thought; and I will suggest that, seen in the long perspective, it is "our time", not "their time", that stands out as exceptional.

1.1 Finding Ourselves in Time

Whatever its ultimate status in the physical universe, there can be no denying that time is a foundational part of the experiential, phenomenal life-world. It is important, however, to try to distinguish this temporal aspect of experience, which we can perhaps assume to be transcultural, from the highly culturally variable conceptualizations of time that we shall explore in following sections. In particular, in describing temporality in experience, we should avoid as far as possible (or at least be cautious about) the use of metaphors, not just of "flow" and "passage" (which imply motion "in time" analogously with motion in space); but also of stative "location in time". The reason for this prescriptive injunction will be become clear: it is not transculturally, we shall see, the case that time is conceptualized as spatial, or as being "like" space.

Time as experienced is made up of the properties of events, which have two basic, perceptible aspects: duration and succession (or sequential order). Duration is temporal extension. Succession is temporal position. In stating this, we are, indeed, immediately inviting, if not relying on, an analogy between duration and spatial extension, and succession and spatial position (in front/behind, before/after). The analogy between temporal succession and spatial order was, of course, the basis for McTaggart's famous distinction between two temporal series, the 'A'-series and the 'B'-series [3]. Events, then, are in some respects like objects; but they are also different. Objects are located in space, and endure, however fleetingly, in time. They have properties like mass and energy. Events are "located" in time, as well as in the space occupied by the objects involved in the event, having properties of duration and succession. Furthermore, we employ temporal (event) landmarks to orient ourselves in time, just as we employ spatial (object) landmarks to orient ourselves in space. However, while spatial landmarks are employed in the service of literal navigation in space, involving physical motion, we cannot physically travel in time, and our temporal navigation is entirely conducted in the mind and in linguistic discourse.

Temporal duration words include adjectives such as "long" and "short", but also measured time intervals such as "ten seconds" and "four months". Temporal landmarks include adverbials such as "today", "yesterday" and "tomorrow", but also named times of day (midnight, three-thirty), dates (22nd June) and other calendrically structured events (Easter, my birthday, Graduation day). We can take as our starting point the plausible-seeming hypothesis that all human beings, transculturally, experience events and inter-event relationships in terms of duration and succession [8]; but that the particular words and concepts denoting temporal duration and temporal landmarks, although they may be based in universal human experiences such as awareness of the diurnal cycle, are based in specific cultural and civilizational traditions, and to that extent are language and culture-specific.

2 Concepts of Time in History and Culture

2.1 The Clock and the Calendar

A striking exemplar of a medieval clock is shown in Fig. 1. Such clocks can be found throughout North-West and Central Europe.

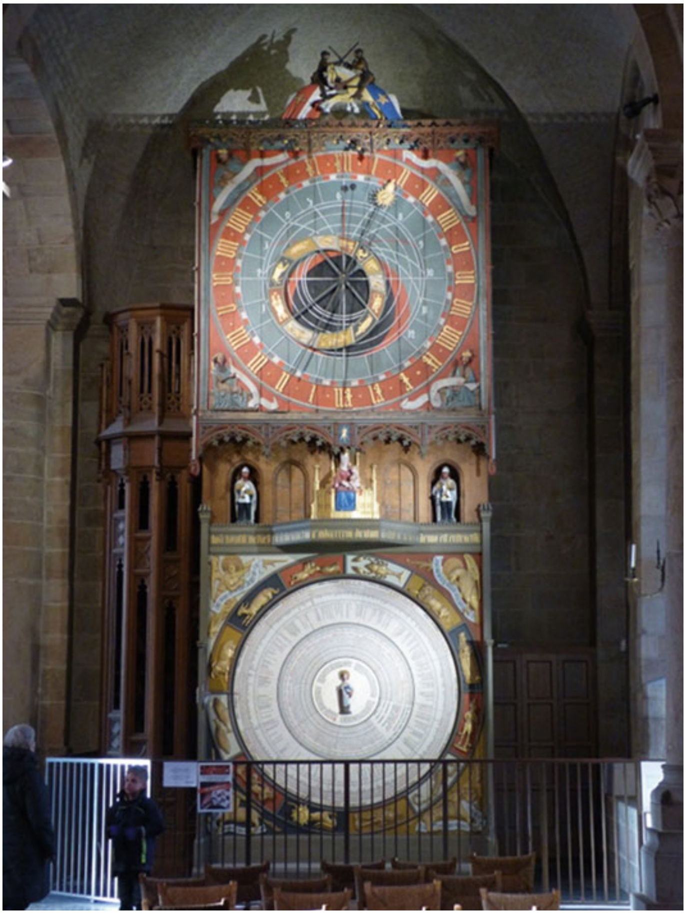


Fig. 1 A medieval clock in Lund Cathedral

Early church and cathedral clocks lacked faces, and sounded the hours by the ringing of bells [9] but later ones incorporated clock faces schematically representing cyclic time intervals—in the case illustrated, not only the hours of the day, but also months and years. The circular form of the clock face iconically represents the cyclic schema which organizes the numerically (ordinally) based time intervals. Although clock hours and calendar intervals are a much older invention than the mechanical clock itself, dating to the Babylonian civilization, these time intervals were dependent upon number notation, as well as upon the astronomical observations measured and notated. Number notations themselves are derived from linguistic number systems whose origins are to be found in counting practices.

It is well known that the cultural dissemination of "calendar time" (which was important in the computation of saints' days), and later "clock time", had profound effects upon medieval and early modern European societies, enabling the accurate determination and registration of both religious festivals and working time [10]. What is perhaps less appreciated is the extent to which the invention and cultural evolution of the calendar and the clock have transformed human cognition, not least by constituting a novel cognitive domain of abstract "Time as Such" [11, 12]. By this, I mean precisely that notion of time, familiar to us as much as it was to Isaac Newton, that metaphorically situates or encompasses the events that occur "in time", and their time of occurrence, analogously to the way that space situates or encompasses objects and their locations.

The universality of concepts or categories of space and time has been a key trope of Euro-American thought since the philosophical reflections of Immanuel Kant [13]. Present-day cognitive science has adopted this hypothesis (although, in many cases, the hypothesis has been more of an unexamined assumption), postulating the existence of a universal cognitive domain (TIME) that (equally universally) recruits its structuring resources from the cognitive domain of SPACE. However, as I have argued above, although both the phenomenological experience of time, and the linguistic encoding of temporal inter-event relationships in lexicon and grammar, may be supposed to be human transcultural universals, the cultural conceptualization and linguistic expression of time intervals (that is, lexicalized concepts of intervals of temporal duration) is widely culturally variable. Much anthropological linguistic research has addressed variability in calendric systems, and in the social practices of "time reckoning" [14, 15] that are dependent on, and realized through, such calendric systems. There has also, however, been another largely unexamined assumption that has informed this research: that however much they may vary, time interval systems are in all cultures cast as some kind of recurrent calendar.

2.2 *Time Interval Systems, "Passage" and Space–Time Metaphor*

Numerically based calendric systems can be regarded as organizing *Time-based time intervals.* Time-based time intervals (such as "Clock Time" and "Calendar Time") are those whose boundaries are constituted by the segmentation and measurement of "Time as Such". Examples of Time-based time intervals are hours and weeks. Although time-based time intervals are based upon natural (astronomical) cycles of events, they are conventional and their duration is derived from counting in a number system.

Time-based time intervals can be distinguished from *Event-based* time intervals. Event-based time intervals are those whose boundaries are constituted by the event itself. In this sense, there is no cognitive differentiation between the time interval and the duration of the event or activity which defines it, and from which in general the lexicalization of the time interval derives. The reference event is often natural (such as 'spring', e.g. "let's take a holiday in the spring"), but sometimes conventional (such as 'coffee break', e.g. "let's discuss this during coffee break"). The event-based time interval may be characterized as a change of state (e.g. 'sunrise'), as a stative event attribute (to use an example from the Amondawa language discussed below, the word *ara* means 'daylight'); or as an activity. The lexicalization may be metonymic or 'pars pro toto', as in Amondawa *pojiwete*, 'when we start work, morning' [16].

Expressions such as "let's take a holiday in the spring", employing locative prepositions to situate one event in temporal relation to another event, are ubiquitous in Indo-European languages. Not only prepositional constructions, but also verbs of motion are employed to conceptualize and express events in time, and their relationship to other events, and the experience of subjects in relation to events. "The summer passed quickly", "your exams are coming up" and "her vacation is approaching" are examples of linguistic constructions in which events "move" along a time line with respect to the phenomenological "now" of the experiencer (the speaker, the addressee or a third party, respectively). A different construction type conceptualizes the experiencer as moving along the time line with respect to static or fixed events, as in: "I left the things of childhood behind", "you are coming up to your exams", "he is past his prime". Constructions of the first type have been called "Moving Time", and of the second "Moving Ego" [17].

Moving Time and Moving Ego constructions are two variants of what we might call the generic, metaphoric "Passage Construction". Passage constructions can be used with reference to either Event-based time intervals ("the summer has gone by") or Time-based time intervals ("the Friday deadline is approaching"). Not only prepositional and other locative constructions for talking about time and temporal relationships, but also Moving Time and Moving Ego constructions have been found to occur in a wide variety of the languages of the world [18]. It has been suggested that this prevalence of using terms and constructions whose primary, more basic meanings relate to spatial location and motion, to express concepts of time and temporal relations, attests to a human cognitive universal. Fauconnier and Turner [19], for example, claim that "Time as Space is a deep metaphor for all human beings. It is common across cultures, psychologically real, productive and profoundly entrenched in thought and language". It is this strong universalist claim that Whorf anticipated and challenged, at least with respect to the "Passage" metaphor and its linguistic expression, in his analysis of the Hopi language. It is a claim that is also thrown into question in the light of research my colleagues and I carried out on concepts of time, and the language of space and time, in the culture and language of the Amondawa, an indigenous Amazonian community speaking a Tupi Kawahib language [11, 12].

2.3 Time in the Amondawa Language

Our findings can be summarized as follows. First, we found that the Amondawa language has a rich variety of lexical and grammatical resources for conceptualizing and expressing spatial relations and spatial motion [20]. Amondawa employs a system of locative postpositions with meanings not unlike those of English or French locative prepositions. It expresses spatial motion in a way more akin to Romance languages such as French or Spanish, than to Germanic languages such as English or German, using "path conflating" verbs of motion like *sauter*, rather than generic motion verbs with satellite particles, such as *go out* [21]. Although some interesting features of the Amondawa language led us to propose modifications of previous linguistic typologies of spatial motion, the language presented no characteristics that were radically different from those described for other languages and language families. It certainly could not be maintained that the language of space in Amondawa, and the resources afforded by it for conceptualizing and expressing spatial relations and spatial movement, is in any respect impoverished in comparison with, say, English or Italian.

Our findings regarding the language of time in Amondawa, however, presented a startlingly different picture. Our data suggest that this language presents a counter-example to the often-assumed universality of space-to-time metaphoric mapping. Amondawa speakers who are bilingual in Portuguese, while able to understand space–time metaphoric constructions in Portuguese, insist that such constructions do not exist in Amondawa, even though the equivalent spatial motion constructions exist [11]. We established in our research that the non-existence in Amondawa of space–time metaphoric constructions is not a consequence of their being ungrammatical; nor is it a consequence of a generalized lack of metaphor in the language. Rather, it seems that space–time metaphorical mapping has simply not emerged, or been "invented", in this language. Why might this be the case?

Other findings, relating to time interval concepts in Amondawa, may hold the clue as to why space–time metaphors are absent in the language. The first thing to note is that Amondawa is one of many Amazonian languages that are known to have very restricted number systems. Small number system languages generally lack numerals above four of five; Amondawa is typical of such languages, in having only four numbers, with larger numbers being indicated by lexical and intensifying variations on words meaning “many”. Clearly, a calendar of the kind that we are familiar with, involving weekly, monthly and annual day counts, simply cannot be constructed in a small number language such as Amondawa. Unsurprisingly, therefore, Amondawa lacks a calendric system in which days of the week or months of the year are enumerated.

More surprising, however, is the complete lack in Amondawa of names for the basic lunar and/or solar time interval units that are often considered to be trans-culturally universal: weeks, months and years. Amondawa has no words for any of these time intervals, all of which, when enumerated, are what I called above Time-based time intervals. As far as we could establish in our investigations, the only time interval units based on natural cycles in Amondawa are day, night and the two seasons, dry and rainy. There is no superordinate “year” in Amondawa, composed of a dry season/rainy season combination.

None of this means that the Amondawa life world is one in which time is absent. The language, while it lacks (like many other languages) a verbal tense system, possesses a nominal aspect system that can mark objects as having a certain status belonging to the past or future (rather as in English, for example, we can talk about an “ex-husband” or a “wife-to-be”). Furthermore, events can be designated as occurring in the future or the past relative to a deictic present, similarly to English “yesterday” or “tomorrow”, and “then”; one event can be expressed as being co-temporaneous with another; and narrative sequences can inter-relate events sequentially. Nonetheless, even if the Amondawa have a similar phenomenological *experience* of time as a “passing” of events, and even if for them, as for us, duration is a fundamental quality of experience, it seems that they do not *think about* or *talk about* time in the same way that “we” (inheritors of a millennia-long tradition of time measurement, and of the progressive implementation of the clock and calendar as fundamental regulators of social life) think about it and talk about it.

Perhaps the most important clue to this difference is that there is no word in Amondawa translating or corresponding to English *time*, or Portuguese or Italian *tempo*. Amondawa lacks not only Time-based time intervals, but also an abstract concept of time, or what I have called above “Time as Such”. What I am suggesting is that the absence of enumerable, Time-based time intervals in Amondawa, and the corresponding absence of a calendar, amount to the absence of a symbolic cognitive model [22] that culturally and historically potentiates the invention (or cultural-historical construction) of the cognitive-conceptual domain of “Time as Such”.

The consequence is that the concept of ‘time’ has no lexicalization, and the schematization of the domain of time as motion through an imaginary and metaphorical “space”, either cyclical or linear, a schematization that appears to us natural and self-evident, is absent from the Amondawa repertoire of cultural schemas. In short, the conceptual domain of abstract and reified “Time” is not a

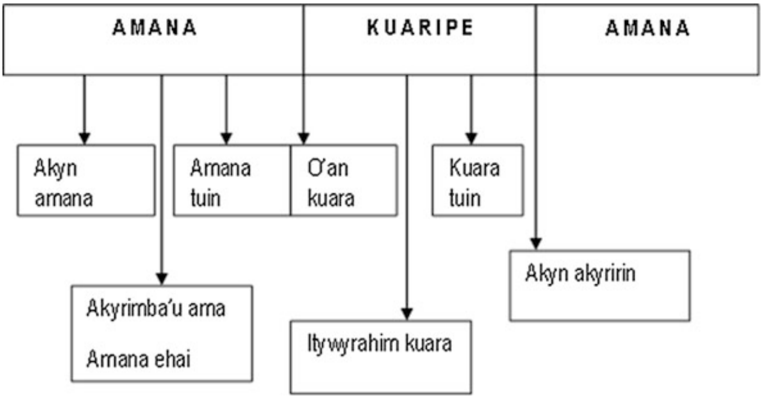


Fig. 2 The Amondawa season schema: a rectilinear depiction of a speaker's sequence

human cognitive universal, but a cultural and historical construction, constituted by schematized time-based time interval systems, reflection upon which is language and culture dependent.

Once again, we should not mistake the absence of a specific mode of domain-constituting schematization for a generalized lack of cultural schemas for time intervals. Amondawa does indeed have a system of seasonal time intervals, with 3 sub-intervals embedded in each of the two superordinate seasons, 'Amana' (*rain* = rainy season) and 'Kuaripé' (*in the sun* = dry season). The subdivisions correspond to the beginning, middle and end of each of these seasons (Fig. 2).

Neither should it be supposed that Amondawa speakers are incapable of, or especially resistant to, making mappings between temporal and spatial schemas. Figure 2 is based upon a spatialization task that we carried out with Amondawa speakers, in which they placed paper plates representing seasonal sub-intervals on the ground [11, 12]. Participants had no difficulty in completing this task, although their representations were curvilinear rather than rectilinear.

Subdivisions of the day are based upon activities that typically and normatively take place at certain times. Although the Amondawa are attentive to the position of the sun at different times of the day, these positions are indicators of socially-structured time intervals, rather than points or positions in a count series. Interestingly, Amondawa participants rejected the circular, or cyclical, depictions of their day/night schema shown in Fig. 3, presented to them by researchers, as being incorrect representations.

It is also interesting to note, from the point of view of indigenous theories of human development, that in the absence of a large number system, the Amondawa do not entertain cardinal chronologies such as ages of individuals. Life stage changes are marked by individuals changing their proper names, in an onomastic system involving the adoption of names from an inventory that marks not only "age" (life stage), but also gender and moiety (sub-clan) affiliation [11, 12].

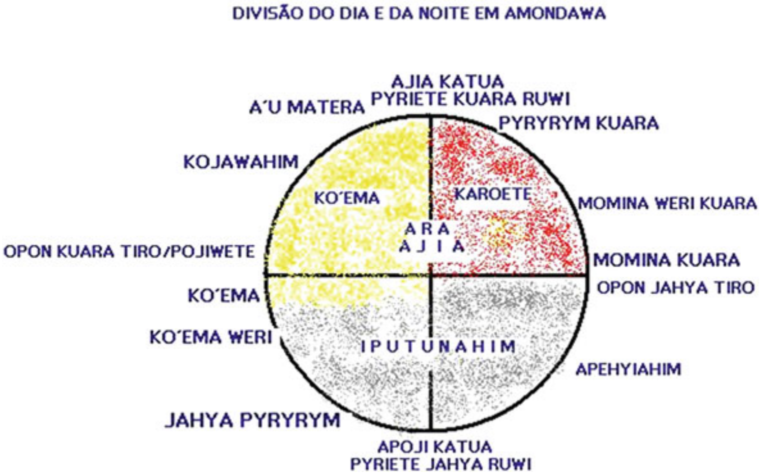


Fig. 3 Divisions and subdivisions of day in night in Amondawa: a cyclical representation rejected by Amondawa speakers

In summary, time, in the Amondawa language and culture, is based not on countable units, but on social activity, kinship and ecological regularity. The absence of artifacts such as clocks and calendars, I suggest, is the motivating reason behind the absence of the cultural-cognitive concept of 'Time as Such'; and it is in the absence of these artifacts that we should also seek the reason for the absence of space–time metaphoric mapping in the Amondawa language.

Although this interpretation of these research findings remains, at this stage, a hypothesis, it is consistent with account Benjamin Lee Whorf's account, cited in Sect. 1, of conceptions of time in the Hopi language and culture. He claimed, recall, that the Hopi speaker "has no general notion or intuition of *time* as a smooth flowing continuum in which everything in the universe proceeds at an equal rate, out of a future, through a present, into a past; or, in which, to reverse the picture, the observer is being carried in the stream of duration continuously away from a past and into a future". In other words, Whorf claimed that Passage (Moving Ego and Moving Time) construals of time were absent in Hopi, just as we claim that they are also absent in Amondawa.

Such effects of language and culture on thought in no way imply an absence of universal cognitive capacities [23]. In fact, our data clearly demonstrate that even when entrenched, habitual, regular linguistic space–time mapping is *absent*, the cognitive capacity for construing temporal concepts in terms of spatial arrays is present in Amondawa speakers; indeed the tasks that we administered *depend upon* the language informants' capacities to make such construals. The explanation we have advanced quite explicitly does *not* propose any generalized absence of the capacity for cognitive space–time mapping on the part of speakers of Amondawa (or any other human group).

In short, the hypothesis my colleagues and I are advancing is that the cognitive domain of "Time as Such" is not a transcultural universal, but a historical construction based in social practice, semiotically mediated by symbolic and cultural-cognitive models for time-based time interval reckoning, and subsequently entrenched in lexico-grammar. Linguistic space–time mapping, and the recruitment of spatial language for structuring temporal relations, is consequent on the cultural construction of this cognitive and linguistic domain. In the consolidation of this constructive process, the invention and perfection of certain kinds of artifacts, clocks and calendars, has been of fundamental importance. Clocks and calendars, I shall now argue, exemplify a more general class of artifact that has been of fundamental importance in the evolution of the human cognitive niche [24, 25].

3 Symbolic Cognitive Artifacts

Language, as many authors have maintained, is grounded in embodied interactional relationships between developing human organisms and their material, social and symbolic surround. Language, and the cultural practices and processes that language supports, are traditionally designated in anthropology and archaeology as symbolic culture, in contradistinction to the ensemble of human artifacts that make up material culture. This dichotomy between material culture and symbolic culture has had the unfortunate consequence that the meaningfulness and social-cognitive agency of the world of material, artifactual objects has been under-investigated [26]. As I shall demonstrate in this section, all artifacts have semiotic, as well as physical, properties. I will go on to argue that language is not only grounded in human interactions with material culture, but is also the symbolic ground of a special subclass of artifacts that I designate **symbolic cognitive artifacts**. This subclass can be defined as comprising those artifacts that support symbolic and conceptual processes in abstract conceptual domains, such as time and number.[1]

Examples of symbolic cognitive artifacts are notational systems (including writing and numeric notations), dials, calendars and compasses. Symbolic and/or cognitive artifacts [31] have been plausibly proposed as key components of human cognitive evolution, in virtue of their status as external representations of cultural and symbolic practices [32], and embodiments of the "ratchet effect" [33] in cultural evolution. While not demurring from this perspective, I will attempt to advance the argument further, by proposing that symbolic cognitive artifacts have the status of agents of change in cultural-cognitive evolution, and are not mere repositories of prior changes in practices and cognitive structures and strategies.

[1] "Symbol" and "symbolic" are notoriously polysemous and contested concepts. In accordance with Karl Bühler's classification [27], symbolicity is here understood in terms of the semiotic, pragmatic and intersubjective logic of communicative representation [28, 29], not on the typology in the Peircian sense [30] of the relationship between sign and object.

Cultural and cognitive schemas organizing at least some conceptual domains (including that of time) may be considered, I shall argue, as *dependent upon*, and not merely expressed by, the employment of symbolic artifacts in cultural and cognitive practices.

3.1 Artifacts, Cognition and Signification

All (human) artifacts are cognitive, inasmuch as they embody human intentionality [28, 34]. However, the semiotic properties of artifacts have received scant attention. In general, artifacts have the following characteristics:

- Artifacts are made, not found. Although found objects may be used as tools, as with for example the sticks that chimpanzees use for "fishing" termites; or as constituents of artifacts, as with stones used by humans to construct dwellings and walls, artifacts (including artifactual tools) are produced by labor.
- Artifacts embody intentionality, conceptualization and imagination. An artifact is made according to a plan or design that involves the conceptual or imaginative representation by the maker of the finished article. It is this characteristic that distinguishes true artifacts from quasi-artifacts, and as far as we know the only species that produces true artifacts is homo sapiens, or as our species has also been aptly named, *homo faber*.
- Artifacts have canonical functions [28] that are physically realized in the design features (or culturally produced affordances) of the artifact. The canonical function of an artifact is equivalent to the use value [35] for which it was designed: its socially-standard function. Non-artifactual (natural) objects or materials (such as wood or stone) may have use-values, but only artifacts have canonical functions. The canonical function of the artifact is embodied in the artifact. For example, the canonical function of a knife is to cut, the canonical function of a cup is to contain. The artifact can therefore be seen as embodying functional or relational concepts, such as CUTTING or CONTAINMENT, and these concepts are precisely those that are the objects of the design intentions of the maker.
- Artifacts signify their canonical function to a user who has the cognitive capacity to recognize the artifact as a token of a particular type [36]. The mode of signification that is intrinsic to the artifact is that of "counting as" [37]. For example, a particular object (token) counts as a cup (type) if the perceiving subject recognizes the design features of the object (being a solid of a certain size and shape, having a cavity affording containment) as being those of a cup. This recognition of the signification relationship of "counting as" is a case of "perceiving as"—the subject perceives the object *as* a cup. If the object is not perceived as a token of a type having a canonical function, then it cannot be said to count as that type for the subject.
- To count as a type of artifact it is necessary for an object not only to afford the canonical function of the type (e.g. containment), but for this to be the

intentionally designed canonical function of the token. For example, a half coconut shell can be used as a cup, but that does not make it a cup, unless it is intended to count as a cup, by virtue either of context or of baptismal naming.

- The counting as relationship, and the canonical function that defines the artifactual type, are normative and cognitive. They are aspects of normative and socially complex cognition. Canonical function depends upon, but is not reducible to, the physical properties of the object, since it is only by virtue of some subset of its physical characteristics (those that enable the object to be perceived as and used as a token of the artifactual type), and of their signifying value for the subject/agent, that the object counts as that artifact. We can thus compare artifacts with "institutional facts" [37], such as that a person is someone else's sister-in-law, a social relationship that is also irreducible to the properties of the person's physical body.

The characteristics listed above make it clear that artifacts are cognitively and semiotically complex. Artifacts (ranging from tools and vessels to notations and images) can be "read" (in the sense of "perceived as"), but (unless they are textual artifacts) they are not texts.[2] The canonical functions that are served by artifacts are diverse, since they may be implicated in a wide range of cultural practices, both sacred and profane, including ritual, ornamentation, representation and narration, as well as technology. Artifacts can support both non-representational practices (such as cutting and sewing) and representational practices (such as drawing and signposting). Although not all artifacts are representational (bear in mind that artifacts do not represent, or *stand for*, their canonical function, rather they *signify* it by *counting as* the type defined by that function), some artifacts (such as pictures and texts) are representational, embodying the semiotic "standing for" function in addition to the counting as function.

My prime concern here is with technological artifacts, that is tools or tool complexes, whose canonical functions involve the amplification of the natural physical and/or mental powers of the agent—"Conceptualization of artifacts is a form of empowerment" [36] p. 311. Technologies may be classified in terms of the different kinds of powers that they amplify: motor (e.g. the hammer); perceptual (e.g. the telescope or telephone); or cognitive (e.g. the abacus). There is also, however, a further dimension in the typology of technological artifacts, namely the dimension of augmentation *vs* constitution of the powers of the agent. Some technologies amplify the powers of the agent by augmenting already existing capacities and practices. For example, a bow and arrow augments the muscular power of the agent, enabling the arrow to be projected further and with a higher velocity than would be possible by throwing. Other technologies amplify the agent's powers by potentiating and constituting entirely new practices. For example, a needle and thread potentiate sewing, a practice that would be impossible without the use of the technology, which can therefore be considered as constitutive of the practice.

[2] This is an important *caveat*, distancing this analysis from post-modernist theories.

The comparison between signs (including the signs of language) and tools has often been made. Karl Bühler [27], influenced by the functionalism of Prague School linguistics, proposed the Organon (Greek = tool or instrument) Model of language. Lev Vygotsky [38] also viewed signs as instruments, not only enabling communication between individuals, but also transforming intra-individual cognition. Vygotsky regarded the analogy as resting on the fact that both sign and tool support mediated activity; but he also distinguished between their modes of mediation in that, while tools are "outer directed", transforming the material world, signs are "inner directed", transforming and governing mind, self and behavior [38] pp. 54–55. Vygotsky emphasized the importance of semiotic mediation in transforming cognition and cognitive development, focusing on the internalization of conventional signs originating in contexts of discursive practice. He attributed great importance to the formative role of language in the emergence of "inner speech" and "verbal thought", but his employment of the concept of semiotic mediation also encompassed the use of non-systematic signs, including objects-as-signifers. He paid little attention, however, to the role of culturally produced, linguistically grounded symbolic cognitive artifacts.

Although I do not wish to advocate a unicausal technological determinist view of history, it is important to note that the socio-cultural consequences of practice-constituting technologies, and combinations of technologies, may be profound. Benedict Anderson [39] discusses the emergence in the 16–17th centuries of what he calls "print capitalism". Mercantile capitalism based upon trade was not new, but the rapid dissemination of information made possible by print media, such as shipping lists and newspapers, potentiated the emergence of the limited stock company, a new institutional form that transformed the world, ushering in the first era of economic globalization.

We might refer here, too, to the rather earlier invention of double-entry book-keeping as an accounting device permitting accurate recording and balancing of profits, losses, liabilities and assets. Double entry book-keeping is a good example of the specific kind of artifact that I have referred to above as a symbolic cognitive artifact, the fundamental form of cognitive technology. Double entry book-keeping is a technique for the ordering of symbolic (numeric) information, in such a way that it permits the checking and auditing of accounts. It is not only desirable for individual traders, but it also provides necessary evidential support for the trust-based interpersonal relations involved in joint financial enterprises. Like other symbolic cognitive artifacts, it is a tool for thought [40] that is transformative of both the individual mind and the shared, intersubjective mind.

To qualify as symbolic, an artifact must have a representational function, in the Bühlerian sense. All artifacts, as I pointed out above, have a signifying status, inasmuch as they functionally "count as" instances of the artifact class of which they are a member, to use Searle's expression [37]; and their material form signifies their canonical function [41, 42]. However, to be a *symbolic* artifact, the artifact must also represent something outside itself, through a symbolic sign function realized or embodied in the artifact. All such sign functions are ultimately grounded in language, although, as we shall see, they also frequently incorporate iconic relations.

The class of symbolic cognitive artifacts can now be defined as comprising those artifacts—which may either be entirely symbolic, such as number systems, or may embed or "anchor" symbolic information in material structures, such as dials [43]—that support symbolic and conceptual processes in abstract conceptual domains. Examples of symbolic cognitive artifacts are notational systems (including writing and number), dials, calendars, clocks and compasses. A key property of symbolic cognitive artifacts is thus that they are both linguistically grounded and conventional. Symbolic cognitive artifacts may be motivated by natural facts, and the human phenomenological experience of these facts, (e.g. the orbit of sun or moon; the number of fingers on a human hand), but they are not determined by them (witness, for example, the variety of arithmetical bases for number systems).

To return to Vygotsky's distinction, symbolic cognitive artifacts are both "outer" (or world) and "inner" (or mind) directed. They are tools that afford and augment human interactions with the natural and social world; and they are simultaneously signs that mediate those interactions (Fig. 4). Intentionally designed symbolic cognitive artifacts, just as much as language, are constitutive parts of the human cognitive niche, and are of fundamental importance in human cultural-cognitive evolution. They are special instances of the *extended embodiment of cognition* [44]. The symbolic systems and conceptual schemas that they support permit the socio-cognitive practices (and the reproduction of these practices through inter-generational transmission) constituting a segment of the life world of individual and group [45]. The invention and use of symbolic cognitive artifacts is a crucial (and species-specific) aspect of the "ratchet effect" [33] in human cultural evolution.

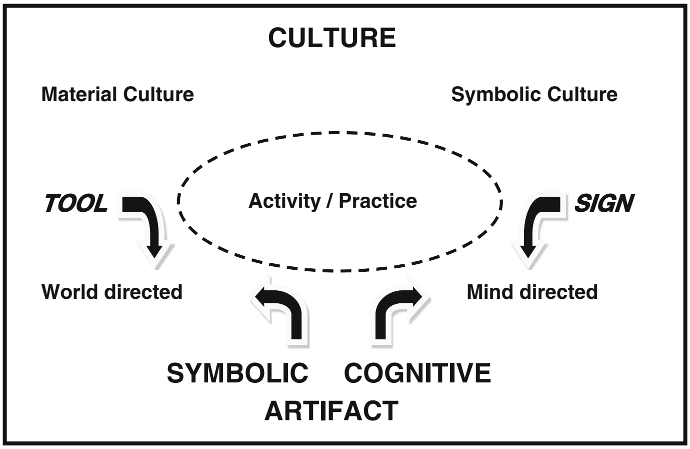


Fig. 4 The bi-directionality of mediated action employing symbolic cognitive artifacts

3.2 Language as a Biocultural Niche and the Evolving Human Semiosphere

Language is the primary and most distinctive constituent of what the Russian semiotician Yuri Lotman called the “semiosphere” [46]: the universe of signs. Signs, as we have seen, are both transformative cognitive tools, and constitutive of specifically human cultural ecologies. The semiosphere can also be viewed, from the perspective of niche construction theory [47], as the semiotic dimension of the human biocultural complex. The self-constructed human biocultural complex both favoured, in prehistory, the emergence and elaboration of language [48]; and, because language is co-constitutive of that niche itself, was fundamentally transformed by language into a symbolic biocultural complex, or semiosphere, introducing a fundamental discontinuity with non-human cultures. This discontinuity has been amplified by the consolidation, through language, of human culture as a fundamentally symbolic order.

Language as a biocultural niche is developmentally and processually interdependent with the “technosphere” of material artefactual supports for acting and for learning through social interaction and social practice. The human semiotic capacity, in collaborative synergy with human constructive praxis, has been the fundamental driving force, in the prehistoric and historical time scale of sociogenesis, of the evolution of human culture and extended human embodiment through the synergistic interplay of semiosphere with technosphere. As Merleau-Ponty put it, “The body is our general medium for having a world … Sometimes the meaning aimed at cannot be achieved by the body’s natural means; it must then build itself an instrument, and it projects thereby around itself a cultural world” [49], p. 146.

This does not imply any need to postulate universal, pre-determined evolutionary pathways. Rather, we need to situate language and cognition in the social ecology of what Pierre Bourdieu called *habitus*: “a subjective but not individual system of internalized structures, schemes of perception, conception and action common to all members of the same group.” A crucial component of *habitus* is constituted by the cognitive symbolic artifacts which, both *grounded in* and *grounding* language structure and language use, serve to develop and expand the uniquely human semiotic biocultural complex [50], p. 86.

4 Living in Time: Inhabiting and Co-habiting Habitus

In speaking and thinking about model-based thinking and reasoning, we tend naturally to adopt a kind of mediated intentional stance: we try to understand how understanding the model will help us to understand the domain that is modeled, that is, the domain that the model is “about”. This domain, we suppose, is in most cases and to varying extents mind-independent, or at least model-independent. Whatever the criteria of adequacy we favor—correspondence, or pragmatic optimization, or coherence with other models—we tend to think of adequacy as approximation to a best fit between model and domain, and of ourselves as independent arbiters of this best fit, standing outside the model-domain relation.

In the case of the time interval systems of clock and calendar that I have addressed in this article, these assumptions simply do not hold. The time that we inhabit is an artifact, a fiction in a way, which is itself the product of the artifacts that our ancestors have invented. Time, we might say, is a cognitive meta-niche, a necessary regulative order for the reproduction of the multiplicity of other cognitive-cultural-material niches that support our activities, practices, communications and reflections. But it is also a cognitive construct, assembled through the spatialization and reification of temporal experience. As Newton pointed out, our secular time-based time interval systems are themselves, ultimately, event-based: clock time and calendar time are derived from the actual motion of celestial bodies. However, when employed to regulate social and economic life, clock and calendar impose a fictive and conventional structure on mundane, terrestrial event time, "freezing" temporal passage into regimes of activity-mapping and time-planning.

The reifying fiction of "Time as Such" is further entrenched in linguistic structure, in "Passage" constructions, and idiomatic usage, in which "time is money", "time is scarce", people are time-poor, and time endlessly presses up against us. The symbolic cognitive artifacts of clock and calendar have changed our minds along with the niches our minds inhabit, and there is no going back in time. And yet, a moment's reflection will tell us that the event-based habitus of the Amondawa, however strange it seems to us when we first encounter it, is the one that has formed the matrix of temporal experience for most human societies. Human beings have lived in small-scale, face-to-face, technologically simple societies for most of the history and prehistory of our species. It is our fast-tracked, globalized, 24/7 turbo-capitalist society that is the exception; and it is we who live by, and have internalized, its insistent imperatives and mind-forged deadlines who are the real (speed-) freaks. Artifactual Time as Such has colonized the niche, and the niche in turn has colonized our minds.

Is that last sentence, too, just a metaphor? If symbolic cognitive artifacts have the effect (as I have argued) of changing both world and mind, is it enough to think of them as mere "tools" for the realization of human deliberative intention, or are they themselves agents? Many discussions of distributed and extended cognition focus on the effects of artifacts on cultural evolution in terms of the externalization of information storage, and the enhanced accuracy of transmission of knowledge and social memory. I would argue that this, while important, is not the whole story. Symbolic cognitive artifacts are not just repositories, they are also agents of change, constituting new domains and potentiating new practices. We can acknowledge that the agency of artifacts is ultimately dependent on human agency, without which artifactual agency would neither exist nor have effect; but it would be wrong to think of artifactual agency as merely derivative, as being like a kind of glorified transmission-belt for human agentive intention. Artifactual agency, I suggest, at least in some cases, is *co-agency*. Co-agentive artefacts play an ever-expanding role in the human biocultural niche, and this poses a real challenge both to our understanding of the nature of knowledge and to our understanding of the nature of ethical and social responsibility in science.

References

1. Newton, I.S.: Philosophiæ Naturalis Principia Mathematica. Royal Society, London (1686)
2. Einstein, A.: Relativity: The Special and the General Theory. Methuen, London (1920)
3. McTaggart, J.E.: The unreality of time. Mind Q Rev. Psychol. Philos. **17**, 456–473 (1908)
4. Smart, J.J.C.: The river of time. Mind New Ser. **58**(232), 483–494 (1949)
5. Whorf, B.L.: Science and linguistics. Technol. Rev. **42**(6), 229–231, 247–248 (1940)
6. Whorf, B.L.: An American Indian model of the universe. Int. J. Am. Linguist. **16**, 67–72 (1950)
7. Minkovsky, H.: Space and time. In: Smart, JJC. (ed.) Problems of Space and Time, vol. 6760, p. 927. Macmillan Pub Co, London (1964)
8. Bergson, H.: Time and Free Will: An Essay on the Immediate Data of Consciousness (Transl. F.L. Pogson). Macmillan, London (1910)
9. Whitrow, G.J.: Time in History: Views of Time from Prehistory to the Present Day. Oxford University Press, Oxford (1988)
10. Postill, J.: Clock and calendar time: a missing anthropological problem. Time Soc. **11**(2/3), 251–270 (2002)
11. Sinha, C., da Silva Sinha, V., Zinken, J., Sampaio, W.: When time is not space: the social and linguistic construction of time intervals and temporal event relations in an Amazonian culture. Lang. Cogn. **31**, 137–169 (2011)
12. da Silva Sinha, V., Sinha, C., Sampaio, W., Zinken, J.: Event-based time intervals in an Amazonian culture. In: Filipović, L., Jaszczolt, K. (eds.) Space and Time in Languages and Cultures 2: Language, Culture, and Cognition, pp. 15–35. Human Cognitive Processing Series 37. John Benjamins, Amsterdam (2012)
13. Kant, I.: Critique of Pure Reason (N. K. Smith, trans.). Palgrave Macmillan, Basingstoke (1929 [1787])
14. Evans-Pritchard, E.E.: Nuer time-reckoning. Afr. J. Int. Afr. Inst. **12**(2), pp. 189–216 (1939)
15. Evans-Pritchard, E.E.: Nuer. Oxford University Press, Oxford (1940)
16. Whitrow, G.J.: Time in History: Views of Time from Prehistory to the Present Day, p. 15. Oxford University Press, Oxford (1988)
17. Clark, H.H.: Space, time, semantics and the child. In: Moore, T.E. (ed.) Cognitive Development and the Acquisition of Language, pp. 27–63. Academic Press, New York (1973)
18. Haspelmath, M.: From Space to Time: Temporal Adverbials in the World's Languages (Lincom Studies in Theoretical Linguistics 3). Lincom Europa, Munich (1997)
19. Fauconnier, G., Turner, M.: Rethinking metaphor. In: Gibbs, R. (ed.) The Cambridge Handbook of Metaphor and Thought, pp. 53–66. Cambridge University Press, Cambridge (2008)
20. Sampaio, W., Sinha, C., da Silva Sinha, V.: Mixing and mapping: motion, path and manner in Amondawa. In Guo, J., Lieven, E., Budwig, N., Ervin-Tripp, S., Nakamura, K., Özçalışkan Ş. (eds.) Crosslinguistic Approaches to the Study of Language. Research in the Tradition of Dan Isaac Slobin, pp. 649–668. Psychology Press, London & New York (2009)
21. Talmy, L.: Lexicalization patterns: semantic structure in lexical forms. In: Shopen, T. (ed.) Language Typology and Syntactic Description. Grammatical Categories and the Lexicon, vol. 3, pp. 36–149. Cambridge University Press, Cambridge (1985)

22. Shore, B.: Culture in Mind: Cognition, Culture and the Problem of Meaning. Oxford University Press, New York (1996)
23. Casasanto, D.: Space for thinking. In: Evans, V., Chilton, P. (eds.) Language, Cognition and Space: The State of the Art and New Directions, pp. 453–478. Equinox, London (2010)
24. Clark, A.: Language, embodiment and the cognitive niche. Trends Cogn. Sci. **10**, 370–374 (2006)
25. Magnani, L.: Abductive Cognition: The Epistemological and Eco-cognitive Dimensions of Hypothetical Reasoning. Cognitive Systems Monographs, vol. 3. Springer, Berlin (2009)
26. Boivin, N.: Material Cultures, Material Minds: The Role of Things in Human Thought, Society and Evolution. Cambridge University Press, Cambridge (2008)
27. Bühler, K.: Theory of Language: The Representational Function of Language (Donald F. Goodwin, trans.). John Benjamins, Amsterdam (1990 [1934])
28. Sinha, C.: Language and Representation: A Socio-Naturalistic Approach to Human Development. Harvester-Wheatsheaf, Hemel Hempstead (1988)
29. Sinha, C.: The evolution of language: from signals to symbols to system. In: Kimbrough Oller, D., Griebel, U. (eds.) Evolution of Communication Systems: A Comparative Approach. Vienna Series in Theoretical Biology, pp. 217–235. MIT Press, Cambridge (2004)
30. Peirce, C.S.: Collected Papers of Charles Sanders Peirce, vols. 1–6, 1931–1935, eds. Charles Hartshorne and Paul Weiss, vols. 7–8, 1958, ed. Arthur W. Burks. Harvard University Press, Cambridge (1931–1958)
31. Norman, D.: Things that Make us Smart. Addison Wesley, Reading (1993)
32. Donald, M.: Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition. Harvard University Press, Cambridge (1991)
33. Tomasello, M.: The Cultural Origins of Human Cognition. Harvard University Press, Cambridge (1999)
34. Bloom, P.: Intention, history, and artifact concepts. Cognition **60**, 1–29 (1996)
35. Marx, K.: Capital, vol. 1. Penguin, Harmondsworth (1976 [1867])
36. Tummolini, L., Castelfranchi, C.: The cognitive and behavioral mediation of institutions: towards an account of institutional actions. Cogn. Syst. Res. **7**(2–3), 307–323 (2006)
37. Searle, J.: The Construction of Social Reality. Penguin, London (1995)
38. Vygotsky, L.S.: Mind in Society: The Development of Higher Mental Processes. Harvard University Press, Cambridge (1978[1930])
39. Anderson, B.: Imagined Communities. Verso, London (1991)
40. Waddington, C.: Tools for Thought. Paladin, St. Albans (1977)
41. Sinha, C., Rodríguez, C.: Language and the signifying object. In: Zlatev, J., Racine, T., Sinha, C., Itkonen, E. (eds.) The Shared Mind: Perspectives on Intersubjectivity, pp. 358–378. John Benjamins, Amsterdam (2008)
42. Sinha, C.: Objects in a storied world: materiality, narrativity and normativity. J. Conscious. Stud. **16**(6–8), 167–190 (2009)
43. Hutchins, E.: Material anchors for conceptual blends. J. Pragmatics **37**, 1555–1577 (2005)
44. Sinha, C., Jensen de López, K.: Language, culture and the embodiment of spatial cognition. Cogn. Linguist. **11**, 17–41 (2000)
45. Schutz, A.: Collected Papers III Studies in Phenomenological Philosophy. Martinus Nijhoff, The Hague (1966)
46. Lotman, Y.: Universe of the Mind: A Semiotic Theory of Culture (Transl. Ann Shukman). I.B. Tauris and Co. Ltd, New York (1990)
47. Odling-Smee, F.J., Laland, K.N., Feldman, M.W.: Niche Construction: The Neglected Process in Evolution. Princeton University Press, Oxford (2003)
48. Odling-Smee, J., Laland, K.N.: Cultural niche-construction: evolution's cradle of language. In: Botha, R., Knight, C. (eds.) The Prehistory of Language, pp. 99–121. Oxford University Press, Oxford (2009)
49. Merleau-Ponty, M.: Phenomenology of Perception. Routledge & Kegan Paul, London (1962)
50. Bourdieu, P.: Outline of a Theory of Practice (Transl. R. Nice). Cambridge University Press, Cambridge (1977)

How Scientific Models Differ from Works of Fiction

Demetris Portides

Abstract Despite the fact that scientific models and works of fiction are the products of the same tools, i.e. idealization or approximation, models differ from works of fiction because the idealization practices in science act as guiding instruments that lead to new knowledge. This epistemic function is characteristic of scientific models and it is a necessary condition for the success of a model, whether theory-driven or phenomenological. It is not a necessary condition for works of fiction because, when these works represent, they represent only the general features of their target thus shaping or improving our intuitions about those features of the respective targets; they do not necessarily lead to knowledge concerning the specific features of the target. Hence, if we treat scientific models as works of fiction in order to understand what models are and how they function, we run the risk of overlooking this distinct kind of epistemic function that models display.

1 Introduction

A growing number of philosophers adopt the view that scientific models are much related to works of fiction; some use this observation to support the view that to understand what models are and how they function we would gain much by treating them as works of fiction (e.g. [7, 10]). Although these views range within a wide spectrum, in this paper I am not concerned with particular versions of how and to what extent scientific models and works of fiction relate. I am mostly interested in supporting the more general thesis that in trying to understand the concept of scientific model by appealing to some of its similarities with works of

D. Portides (✉)
University of Cyprus, Nicosia, Cyprus
e-mail: portides@ucy.ac.cy

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_5, © Springer-Verlag Berlin Heidelberg 2014

fiction, we run the risk of obscuring significant aspects of scientific modeling, and thus this comparison may not be as fruitful as some would claim. This is not to deny that some conceptual elements of models, strictly speaking, represent fictional entities or fictional states of affairs. This may be true, but it has little to do with the purpose of constructing models and with how models function. Furthermore, to infer that models are fictions from the premise that some of their constitutive elements are fictional obscures the epistemic role of models by identifying the aim of scientific discourse with the ideal of literal and complete truth. The latter, however, seems to be unrelated to scientific discourse and to scientific modeling in particular. Although I do not think there exists a school of thought in the philosophy of science which could without hesitation be labeled 'fictionalism' (as we encounter in other areas, e.g. the philosophy of mathematics, ethics), I will use the term to refer to some kind of general objection to the idea that models and works of fiction are distinct in significant ways.

Generally speaking, we label *X* fictional in contrast to labeling it truthful, and since truth is a valuation term for statements or propositions, to label *X* fictional, *X* must be the kind of construct used to convey a statement or proposition. Thus, we label *X* fictional in order to accentuate the fact that the claim made by *X* is in conflict with what we observe the state of the world to be. For instance, we do not label a poem fictional if we simply conceive it as being a linguistic construct that is meant solely for aesthetic purposes. Obviously, sentences are used to make statements, but we can also assume that indirectly models, pictures or analogical representations can be used in different ways to convey a statement. Therefore, more generally, if we allow *X* to be instantiated by any of these types of constructs, then we claim that construct *X*, of whatever instance, is fictional if we think that the truth valuation of the claim "*X* represents (an aspect of) the world" is false. So in labeling *X* fictional or truthful, what is involved is our criteria of *X*'s truth valuation, and not the way(s) by which we construct *X*; and *prima facie* it seems reasonable to contend that the ways by which we construct *X* do not determine in any way the criteria of *X's* truth valuation. This is something I dispute in this paper by developing much of the argument in support for the following idea: what epistemically distinguishes models from works of fiction is how the details of the (simplifying) processes involved in constructing a model influence what must be done in order for the model to accurately represent its target system and thus these processes determine how modeling a target usually proceeds in science.

I should, however, acknowledge from the start that the way(s) by which we construct *X* is similar, in many ways the same, whether *X* is meant for aesthetic or epistemic purposes, or merely for the purpose of expressing and thus rendering a thought or a feeling communal, or for other purposes. In general, these ways involve simplification of one sort or another of the complexities present in the actual world. Normally we think of the simplifications involved in speaking about the world as the result of idealizing or approximating individual elements, followed by a synthesis of the simplified elements, in a description, in ways the individual imagination permits.[1] In short, as Giere [9] correctly claims, for whichever purpose *X* is constructed (e.g. epistemic, aesthetic or other), it is a product of the imagination. As such, a large portion of the statements we construct are strictly speaking false, i.e. if we exclude atomic sentences of the form "it is snowing now in that part of the world", most other statements are false because they involve simplifications. If our criterion for *X*'s truth-value is "its literal truth", then most sentences we utter are false since they describe real-world situations in simplified ways.

Thus, if fictionalism relies on the position that scientific models are fictions because they contain elements that fail to refer to real entities or physical states exactly as they are experienced in all their complexity in the world, then most of what I have to say in this paper leaves such a fictionalist attitude unaffected. If, however, fictionalism relies on the more deflated view that some elements of these models are false (i.e. they fail to refer) *because* they are the product of idealization or approximation (of course, it may be that most of our discourse is fictional because of this), then such a fictionalist attitude can be charged, from the perspective I take in this paper, with partially obscuring aspects of modeling. In particular this fictionalist attitude seems to overlook that despite the fact that scientific models contain 'fictional' elements—which result from idealizing and approximating—they do allow us to systematically explore their respective targets in ways no other kind of discourse—which also results from idealizing and approximating—does. Thus one should explore why this is so, and a fictionalist attitude, in my view, obscures the necessity for doing it. Another way of putting this point is that it is one thing to claim that, because of underlying idealizations or approximations, our scientific representations are not utterly realistic, and it is another thing to claim that they are fictional. Characterizing our representations as fictional restricts us from posing the right questions with regards to why from not outright realistic representations we are led to new knowledge.

Of course, if the above fictionalist positions are underpinned on the view that scientific statements that derive from models are strictly speaking false and that this is enough reason to ignore the success of these models (and also why these particular 'fictions' are so successful as opposed to other kinds of fiction), then to back fictionalism one must invoke arguments in support of the position that "inferring the literal truth of models from their success is a fallacy". I think, however, that to endorse fictionalism about models on the grounds of an argument

[1] I use the term 'idealization' in this paper as a generic term to refer to simplifications that result from the omission of features and to simplifications that result from modifications of the actual features of objects or of states of affairs. Some authors refer to the former as 'abstraction' and to the latter as 'idealization' and maintain that they are products of distinct thought processes, e.g. Cartwright [3]. Of course, there are important differences between the two kinds of simplification that have been pointed out by many authors. However, it is debatable whether these two kinds result from distinct thought processes; see for instance McMullin [12], who treats them as if they are the product of the same cognitive act. Whatever the outcome of such debate, the differences between the two kinds of simplification are largely irrelevant for the purposes of this work.

that demands the ideal of 'literal truth' overlooks other important epistemic functions of models. Teller [20] argues that insisting that our scientific constructs must meet the strict requirement of truth, i.e. having complete precision and accuracy, in order to count as nonfictional is wrong. That this is so, Teller argues, follows from the fact that "… the respects in which a representation counts as veridical [i.e. true enough] will never be respects in which the representation counts as fictional" [20]. My argument has a corollary along the lines of Teller's argument, which is that, because models allow partial epistemic access to their targets that would not result otherwise, the ideal of literal truth is an exaggerated criterion by which to understand this function of models. In other words, because the tools by which models are constructed are idealization or approximation, models are bound to be inexact and inaccurate. However, successful representational models should be considered true enough not just because their predictions approximate the values of experimental measurements or because they bear similarity relations with what they represent, but also because they produce new knowledge about their target.

Here is a sketch of my argument: Scientific models and works of fiction are constructed with the same cognitive tools, i.e. idealization or approximation. The ways idealization (or approximation) is employed in successful scientific models is such that it leads to knowledge about the general features of the targets, but it also guides the users of the model to new knowledge concerning the specific features of the target systems. This is the result of the interplay between the model and the target of the model, and it is a necessary characteristic for a model to be successful in science. Works of fiction can also provide knowledge of the general features of their target. It may even be the case that, some works of fiction are such that they also lead to knowledge of the specific features of their targets. This, however, is not a necessary characteristic of works of fiction, because the possibility for interplay between a work of fiction and its target is minimal, if it can take place. Hence, we cannot use fiction to learn about models without obscuring the above epistemic function of models; it might be possible, however, to use models to learn about particular epistemic functions of particular kinds of works of fiction.

2 Different Purposes for Using Models as Epistemic Agents

Scientific models come in various guises and are built for various purposes. Two important purposes for which they are constructed are (i) to facilitate theory formation and (ii) to apply theory to particular physical systems. In using antecedent knowledge to build models that are useful in forming new theories, we come across cases in the history of science in which such models are transparently fictitious about the targets of the new theory. One such example is the case of Maxwell's fluid vortex model enhanced with particles that operate like idle wheels. But, as Morrison [14] and also Nersessian [17] argue, the purpose of such models is not to represent the target in any realistic way, but to help scientists intuit possible ways by which the variables of the new theory could relate. In the case of Maxwell's model, some form of analogy between elements in the fluid vortex model and elements in electromagnetic field theory points to a similarity in the relational structure, i.e. between the relations of the model's variables and the relations of the variables of the theory, and thus helps to visualize some characteristics of the new theory. The model performs a purely heuristic function, but despite this it has obvious epistemic consequences that are well highlighted in both Morrison's [14] and Nersessian's [17] analyses.

It is, however, important to distinguish Maxwell's fluid vortex model from models that are built in order to apply a particular theory to the phenomena in its domain. The liquid drop model of the nucleus, to which I will return in the next section, is meant to apply Quantum Mechanics to explain the structure of the nucleus. The purpose of the model is such that physicists are not content with just using the model to improve their intuition about the nucleus. When they construct such models as the liquid drop they aspire to make them both explanatory of their targets and acceptably good predictive devices, but they also want them to facilitate investigation into their target that would lead to new knowledge. If these functions are demonstrated by a model that is meant to apply theory to phenomena in its domain, then it is considered successful.

This distinction concerning the purposes for which models are constructed is necessary for understanding what their epistemic functions are; it is not, however, necessarily helpful for understanding works of fiction and their function. Let me clarify what I mean by this. Works of fiction, particularly the relation they bear to their targets, just like the case of scientific models, are not all of the same kind. Just as we can distinguish between scientific models that are used to apply a particular theory and models that depict relations in one domain and by analogy help us intuit the relations of the variables in a newly formed theoretical domain, we can also distinguish between works of fiction on roughly the same grounds.

For example, Tolstoy's *War and Peace* refers to its target directly; it depicts relations between events in a construct of the imagination which may be interpreted to refer to relations in Russia during the Napoleonic wars in the beginning of the nineteenth century. The domain of the novel is the domain of the history of the events of a specific period in a specific environment of our world. Although the relations depicted in the novel are abstracted away from, or are idealized versions of, or are approximated versions of actual human relations, their reference to their target is direct. They are relations between events in a construct of the imagination that are derived from a particular historical context with the purpose of rendering them didactic to the reader, i.e. to enable the extraction of general lessons about the cruelty of war, about history, about human relations, etc.

On the other hand, in George Orwell's *Animal Farm* a complex of relations between animals is created, these relations are made explicit, and the novel is interpreted to act as a good guide for understanding the totalitarian nature of the state of the Soviet Union and the effects of totalitarianism on human relations. By describing relations and properties in one domain, Orwell guides the reader to make intelligible the relations in a disparate domain. By means of these examples I wish to stress that, in works of fiction we sometimes also use familiar domains, or explicitly analyzed domains, to gain intuition into unexplored domains, and other times we also depict the relations of a domain directly. I have chosen two examples of fiction that patently demonstrate some epistemic functions, namely, making some general principles intelligible. Despite the fact that I do not think these functions are on a par with the epistemic functions of models, as I will argue in the next section, my point is the following:

Although works of fiction, just like models, could also be distinguished on the basis of directly depicting a domain or indirectly depicting it (via metaphor, analogy or otherwise), I do not know how useful this distinction is for fiction and I am extremely hesitant to claim that it is necessarily important for understanding how works of fiction represent their target and how they afford epistemic access to their targets (when they do). On the other hand, it is vital in understanding how models function and specifically how they afford epistemic access to their targets. To understand the epistemic functions of models, it seems to me necessary to understand firstly the purpose for which they are constructed. Maxwell's models function epistemically, but their function is not the same as that of models that are used to apply theory, as I shall show in the next section, because their purpose is different.

3 Idealizations in Models Lead to New Knowledge

In works of fiction any kind of idealization can be employed by authors in order to produce what they consider relevant to the idea they wish to motivate. This is also true in everyday discourse. It is also true that in science it is logically possible to employ idealizations in order to produce any model one wishes. But the relevance of this truism to how successful scientific modeling is done is minor. In science, models that involve idealizations that do not guide towards new knowledge are eventually abandoned. In order for a model to be successful it must be constructed with idealizing assumptions that guide the scientific community eventually to new knowledge. This is a necessary characteristic of scientific modeling. There is hardly any doubt that many authors of fiction with acute observation skills and depth in their intuitive understanding of a particular domain are capable of producing works of fiction that meet the same criterion successful scientific models must meet. This is not, however, a necessary condition for producing works of fiction; it is only contingent on the skills of the author. This I think is an important key to understanding how models differ from works of fiction.

In my introduction I claimed that a fictionalist view obscures the epistemic functions of models. Here I want to highlight some of the epistemic functions that are characteristic of scientific models which are constructed for the purpose of applying theory to phenomena. Clearly, when one puts the emphasis, as I do, on how scientific models differ from works of fiction, rather than on their similarities, the focus shifts to a different set of philosophical questions. Namely, the relevant questions to ask in order to understand how models act as epistemic agents are related to idealization, e.g. what distinguishes idealizing practices in scientific modeling from uses of idealization in other forms of discourse? Addressing such questions is not the subject of this paper; rather, in what follows I simply want to emphasize the view that it is precisely through an understanding of how the function of idealization in modeling differs from its function in works of fiction that we can effectively gain insight into the epistemic function of models.

In this paper my concern is primarily with the second important use of models, that of applying theory to phenomena. In such cases, models are constructed by the use of the general principles of theory for the purpose of representing target systems. Surely, by the strict standard of 'literal truth' the representation relation between model and target is never completely realistic. Whether this is enough reason to lead us to a fictionalist view on models must be, I think, subject to a closer examination of the representation relation and the model's epistemic function.

To fruitfully explore the epistemic functions of models it is helpful to distinguish between two general categories of models: theory-driven and phenomenological models. Theory-driven models are direct descendants of theory.[2] They are typically overly schematic representations of their targets, because they depict very few features of their targets and ignore most them. Examples of this kind from classical physics are the wave equation that is used, among other things, to model a vibrating string, and the linear harmonic oscillator that is used to model the pendulum, the mass-spring system, or the torsion pendulum. The linear harmonic oscillator is a linear second-order differential equation of position with respect to time. If it is to be used to represent the pendulum, then this kind of representation would rely on the assumption that the only influences on the pendulum bob are the force of gravity, which acts uniformly, and the tension on the pendulum string that makes the former act as a linear restoring force. The implication of such an assumption is that the pendulum would oscillate with constant amplitude about an axis indefinitely. Of course, such behavior is never observed in any real pendulum, even in those pendulum apparatuses that are set-up under highly controlled laboratory conditions. How could such a model be epistemically valuable?

The answer to this question relates to the distinction between general and specific representation. In scientific modeling the goal is twofold. Models aim to represent both the general features of their targets and the more specific features or specific manifestations of the general features of the targets. Since the starting point is—most frequently—an abstract theoretical principle, it would be much misguided to expect an application of theory to lead directly to a representation of the specific features of physical systems. For instance, Newton's 2nd law dictates

[2] Some philosophers construe theory-driven models as the members of the class of models with which the theory is identified or by which it is presented, e.g. Giere [8], van Fraassen [21], Da Costa and French [6].

that the sum of the forces on a body equals the product of its mass times its acceleration. This law is too general and abstract to apply directly to real-world systems for many reasons, but the one crucial to my argument is that all the forces that influence a particular body in a particular situation are subject to discovery and not necessarily antecedently available knowledge. Obtaining such knowledge is part of the model's primary functions.

Let us go through things in a bit more detail. When physicists employ the linear harmonic oscillator to represent the pendulum, they ask us to imagine that the pendulum system has the following characteristics: that the cord is mass-less and inextensible, that the oscillations are infinitesimal, that the bob is a point-mass, and that a medium is absent. Of course, employing such idealizing assumptions to model the physical system creates a big gap between the theoretical description and the real system and its characteristics. Doing this, however, meets two other epistemic purposes. Firstly, it allows the use of the linear harmonic oscillator to describe how the force of gravity acts in general on all pendulums, i.e. the linear harmonic oscillator is a model-type of physical systems that are under a linear restoring force. Secondly, the model dictates how one should proceed in order to represent the target more accurately. The idealizing assumptions, which are part of the model, dictate what should be done in order to bring the model closer to the real system. Namely, what forces would act on the pendulum: (i) due to the medium, (ii) due to the cord having mass and extension, (iii) due to the bob having spatial extension and (iv) due to the oscillations being finite. In other words, the model (understood as the differential equation that expresses it in conjunction with the assumptions that have to be met in order for the former to capture how a linear restoring force would act on the target) dictates where to look and what to look for in order to discover the forces acting on the target system, i.e. it provides a kind of orientation for the scientist on where to look for influencing factors. Of course, this is not the end of the scientific pursuit, but the beginning of a long process that often leads to discovering the forces and how these forces should be expressed mathematically in order for the equation of motion to be manageable and at the same time to make acceptably good predictions.

In accounting for how the medium affects the pendulum motion, the analysis by Nelson and Olsson [16] on the details of the physics of the pendulum leads to the following kinds of influences: (1) partial rotation of bob, (2) buoyancy, (3) linear damping for wire, (4) quadratic damping for bob, (5) decay of finite amplitude, (6) added mass to the bob's inertia. Similarly, in accounting for the cord's and bob's mass and extension and for the oscillations being finite, their analysis leads to the following kinds of influences: (1) finite amplitude, (2) finite radius of bob, (3) mass of ring, (4) mass of cap, (5) mass of cap screw, (6) mass of wire, (7) flexibility of wire, (8) double pendulum of support ring and pendulum bob, (9) stretching of wire, (10) motion of support. It is quite obvious that these influences are not conceived a priori; they are the result of interplay between the linear harmonic oscillator (i.e. the initial general representation of the pendulum) and the experimental set-up. Without the model and a clear understanding of the idealizations involved in setting up the initial model, it is possible that not all of these influences would have been discovered, nor of course how they influence the behavior of the system. Admittedly, each of the above sixteen effects is eventually expressed approximately or in an idealized way in the overall force function (see [13, 19]). This means that the final predictions of the detailed model that captures not just the general but also the specific features of the target will not exactly fit the experimental measurements. Therefore, if our criterion for the epistemic value of the model is that of "perfect fit", then this model, as well as any other model, obviously fails.

But the function of the model is apparently different: it is to represent the general features of the type target and to orient and guide the scientist in order for her to intuit and discover the specific features. Thus the model is an epistemic tool that, on the one hand, makes the general features of the target intelligible via the underlying idealizations and, on the other, through interplay with the target, gradually leads to gaining understanding about how to overcome the idealizations involved and furthermore how to represent the specific features of the target. I doubt whether this kind of epistemic function can be performed by works of fiction.

One could object to this argument by claiming that, as a matter of fact, very few models of the theory perform the above function, i.e. very few models can be used to discover specific features of their targets. Therefore, we are not justified in generalizing what we observe to be the case for a small number of scientific models. In principle, it is possible to define an infinite number of theory-driven models. In actual science very few of them are useful in the above sense. I mentioned above two examples of this kind from classical mechanics; from quantum mechanics examples that could be cited are the infinite harmonic oscillator and the infinite square well. In general, these are the models that Cartwright [4] has labeled 'stock' models of the theory. Addressing this kind of objection comes in two steps. The first step is the obvious. The few stock models that are employed in applying each theory are the only ones that occasionally perform the representational function and thus have the epistemic role attributed to them above. The remaining theory-driven models are just epistemically useless. The second step is lengthier; it concerns the role of phenomenological models, particularly their role as epistemic surrogates in cases where a suitable theory-driven model is unavailable for the exploration of a physical system, or in cases where the suitable theory-driven models cannot be used to gain physical insight into the specific workings of the system. In many such cases phenomenological models are the ones that do the job and afford epistemic access into the targets of the theory.[3]

In science, in addition to theory-driven models, we also encounter models that are not direct derivatives of theory and often are, strictly speaking, in conflict with some principles of the theory they are meant to apply. I call phenomenological models those scientific models that are not directly derived from theory and are

[3] In the last few decades, with the advent of powerful computing machines, computer simulation techniques also perform this function.

constructed by means of a conceptual apparatus that is either independent of theory or vaguely related to the theoretical apparatus.[4] An example of this kind is the liquid drop model of the nuclear structure, which is constructed by assuming the nucleus as having classical properties, and, once the Hamiltonian function is set-up, the parameters of position and momentum are quantized, thus resulting in – what Cartwright [2] dubbed— theory entry of the model.[5]

The problem of using theory to directly represent physical systems becomes apparent when we look into the problem faced by physicists in modeling the structure of the nucleus. Of course, in order to model the nucleus one starts from the Schrodinger equation and attempts to set up the proper Hamiltonian operator for representing the target. But in order to achieve this, one must include in the potential operator the pairwise interactions between all the nucleons. This, even if enough knowledge were available to allow for an accurate characterization of the pairwise interactions, would lead to the nuclear many-body problem, i.e. the inability to solve the Schrodinger equation for more than two-particle interaction. Faced with this problem, physicists moved towards constructing the Hamiltonian operator phenomenologically. What this meant, in the particular historical context, was to search for ways to construct models with particular Hamiltonian operators whose primary initial goal was to explain the available experimental results (i.e. the electric quadrupole moments of the nucleus, and nuclear fission) and other semi-empirical results (in particular, the Wiezacker semi-empirical mass formula, for a more detailed discussion of this point see [19]). In other words, because a direct application of the Schrodinger equation leads to intractability problems, physicists employ various intuitively guided ways with the purpose of explaining the known experimental results.[6]

The liquid drop model is one such case. It offers a plausible explanation for the nuclear electric quadrupole moments as well as for nuclear fission, and it affords good predictions for these phenomena at least to a first approximation.[7] That the liquid drop model is what physicists call semi-classical is easy to see without going into the mathematical technicalities of the model. To construct the model, the nuclear properties are assumed to be subject to a strong interaction force, i.e. the nucleus is assumed to consist of strongly-coupled particles such that it does not demonstrate independent nucleon motion. Thus the nucleus as a collection of particles demonstrates collective modes of motion, much like a liquid drop which

[4] See Portides [18] for a more detailed description and a closer examination of phenomenological models.

[5] In the absence of reductive rules of classical functions to quantum mechanical operators, this is, of course, an arbitrary move. Although this is frequently the case in quantum mechanical modeling, it is not an issue of concern in this paper.

[6] I say ‘initial’ goal, because once a model is successful in its initial goal then it is refined, corrected or modified in order to meet other goals that an acceptable representation should. That is to say, the goals of a putative representational model are not necessarily all set in the very beginning but they can be subject to change throughout its evolutionary history.

[7] See Moszkowski [15] for more details regarding the liquid drop model.

is observed to be moving its surface but not its constituent molecules. The Hamiltonian of a strongly-coupled collection of particles is set up as a classical function (much like that of a liquid drop) that represents a rotation mode, a vibration mode and a rotation-vibration mode of motion, then it is quantized and turned into the potential part of the Hamiltonian operator that is used in the Schrodinger equation for modeling the nucleus.

Despite the fact that this model is not a genuine outcome of quantum theory, it is considered to be successful mainly for three reasons. The first has already been noted: the model provides a good explanation for already established experimental results, namely the electric quadrupole moments and nuclear fission. The second is its successful integration into more refined and successful models of the nucleus, such as the unified model of nuclear structure. The third reason is the one I wish to highlight because of its relevance to the argument in this paper. In 1947 high frequency collective excitations of nuclei, which came to be known as giant resonance (see [5]), were experimentally observed in photonuclear reactions by Baldwin and Klaiber [1]. According to Quantum theory, this phenomenon could only be due to a mechanism that imparts energy to the nucleons, for instance when coincident photons (γ-rays) couple to nuclei and add energy to them. At the time, the liquid drop model was the only model which was easily refined to provide an explanation for the phenomenon. A schematic form of the explanation provided by the model is the following. Giant resonance motion involves density fluctuations (i.e. collective oscillations) in the nucleus that may be caused by the electric field of a coincident photon (γ-ray). The electric field of the photon acts only on the charged nucleons (i.e. protons), and because the nuclear centre of mass has to be at rest, the neutrons move in the opposite direction to that of the protons. The refined liquid drop Hamiltonian operator that accommodates the phenomenon of giant resonance was extended to allow for the demonstration of various types of nuclear density fluctuations. For instance, a dipole fluctuation (in an assumed spherical nucleus) involves the motion of the protons in roughly one hemisphere and of the neutrons in the opposite hemisphere, a quadrupole density fluctuation involves the motion of protons in two opposite quadrants of the sphere and of neutrons in the remaining two quadrants, and so forth for higher order fluctuations. Now, the interesting thing is that Baldwin and Klaiber observed the giant dipole resonance in 1947. As a consequence the Hamiltonian of the liquid drop was refined to accommodate an explanation for this observation. However, the Hamiltonian was designed in such a way that it could be used as the basis for predicting other possible forms of giant resonance. What followed was the observation of giant quadrupole resonance in 1972 and the observation of giant monopole resonance in 1977 (see for instance Harakeh et al. [11]. Both of these phenomena were predicted and explained well by the refined liquid drop, so it can be argued that they belong to the set of novel observations that the model led to.

My point, just as for the case of the linear harmonic oscillator, is that the liquid drop acted as a guide to discovering new phenomena and new properties of its target. I think the reason for the model to act so is its underlying idealizing assumption. By constructing the Hamiltonian operator on the analogy of a liquid drop, one of the primary underlying assumptions is that the nucleus is under the influence of a spherical potential. Of course, this is an idealization and physicists know that. But precisely this idealization is the guiding force behind the drive to discover novel nuclear phenomena: any observations that could imply deviations from sphericity have to be explained and the mechanisms responsible for these deviations must be accounted for in a refined model. The process is a continuous attempt to refine the model such that the initial idealized model through a series of refinements, subject to novel experimental evidence, gradually converges to a more (approximately) realistic representation of its target. This is the result of interplay between the model and measurements on the different behaviors of its target. This converging process is common to both theory-driven models that initially represent the general features of their targets and successful phenomenological models.

4 Conclusion

I have argued that focusing on the similarities between works of fiction and scientific models can lead to overlooking the characteristics of the epistemic functions of models. I chose to focus on the differences to avoid such an outcome. One of the aspects of model building on which light is shed from this perspective is that the idealizing assumptions that underlie a model themselves act as guiding instruments by which new knowledge can be produced from the model. This epistemic function is characteristic of scientific models and it is a necessary condition for the success of a model, whether theory-driven or phenomenological. It is not a necessary feature of works of fiction, because works of fiction represent—when their purpose is to do so—only the general features of their target and facilitate the improvement of our intuitions into those features of the respective targets. They are not intended as guiding instruments for discovering the specific features of the world, and even if they were, they are not capable of systematically doing the job. Scientific models are. This argument reveals, I think, three significant epistemic differences between scientific models and works of fiction. First, scientific models necessarily have an epistemic function, whereas works of fiction do not. Second, even where works of fiction do have an explicit or implicit epistemic function, e.g. Tolstoy's *War and Peace*, works of fiction only aim to represent the general features of their target. On the other hand, models aim also to represent the specific features of their target, even when this is impeded by the absence of the necessary skills and knowledge. Third, even if works of fiction were aiming at representing specific features of their targets, they could not systematically deliver. Scientific models do systematically deliver representations of specific features of the world, albeit in idealized and approximate ways.

References

1. Baldwin, G.C., Klaiber, G.: Photo-fission in heavy elements. Phys. Rev. **71**(1), 3–10 (1947)
2. Cartwright, N.D.: How the Laws of Physics Lie. Clarendon Press, Oxford (1983)
3. Cartwright, N.D.: Nature's Capacities and Their Measurement. Clarendon Press, Oxford (1989)
4. Cartwright, N.D.: Models and the limits of theory: quantum Hamiltonians and the BCS models of superconductivity. In: Morgan, M.S., Morrison, M. (eds.) Models as Mediators: Perspectives on Natural and Social Science, pp. 241–281. Cambridge University Press, Cambridge (1999)
5. Chomaz, P.: Collective excitations in nuclei. http://www-ist.cea.fr/publicea/exl-doc/00000033058.pdf (1997)
6. Da Costa, N.C.A., French, S.: Science and Partial Truth. Oxford University Press, Oxford (2003)
7. Fine, A.: Fictionalism. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modeling and Idealization, pp. 19–36. Routledge, New York (2009)
8. Giere, R.: Explaining Science: A Cognitive Approach. The University of Chicago Press, Chicago (1989)
9. Giere, R.: Why scientific models should not be regarded as works of fiction. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modeling and Idealization, pp. 248–258. Routledge, New York (2009)
10. Godfrey-Smith, P.: The strategy of model-based science. Biol. Philos. **21**, 725–740 (2006)
11. Harakeh, M.N., Van der Borg, K., Ishimatsu, T., Morsch, H.P., Van der Woude, A., Bertrand, F.E.: Direct evidence for a new giant resonance at 80 A–1/3 MeV in the lead region. Phys. Rev. Lett. **38**(13), 676–679 (1977)
12. McMullin, E.: Galilean idealisation. Stud. Hist. Philos. Sci. **16**, 247–273 (1985)
13. Morrison, M.: Models as autonomous agents. In: Morgan, M.S., Morrison, M. (eds.) Models as Mediators: Perspectives on Natural and Social Science, pp. 38–65. Cambridge University Press, Cambridge (1999)
14. Morrison, M.: Fictions, representations, and reality. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modeling and Idealization, pp. 110–135. Routledge, New York (2009)
15. Moszkowski, S.A.: Models of Nuclear Structure. In: Flügge, S. (ed.) Encyclopedia of Physics: Structure of Atomic Nuclei, vol. 39 pp. 411–550. Springer Verlag, Berlin (1957)
16. Nelson, R.A., Olsson, M.G.: The Pendulum—Rich Physics from a Simple System. Am. J. Phys. **54**(2), 112–121 (1986)
17. Nersessian, N: Abstraction via generic modeling in concept formation in science, In: Jones, M., Cartwright, N. (eds.) Idealization XII: Correcting the Model. Poznan Studies in Philosophy of the Sciences and the Humanities, pp. 117–143. Rodopi, Amsterdam (2005)
18. Portides, D.: Seeking representations of phenomena: phenomenological models. Stud. Hist. Philos. Sci. **42**, 334–341 (2011)
19. Portides, D.: Scientific representation, denotation and explanatory power. In: Raftopoulos, A., Machamer, P. (eds.) Perception, Realism and the Problem of Reference, pp. 239–261. Cambridge University Press, Cambridge (2012)
20. Teller, P.: Fictions, fictionalization, and truth in science. In: Suarez, M. (ed.) Fictions in Science: Philosophical Essays on Modeling and Idealization, pp. 235–247. Routledge, New York (2009)
21. Van Fraassen, B.C.: The Scientific Image. Clarendon Press, Oxford (1980)

What’s in a Diagram?

On the Classification of Symbols, Figures and Diagrams

Mikkel Willum Johansen

Abstract In this paper I analyze the cognitive function of symbols, figures and diagrams. The analysis shows that although all three representational forms serve to externalize mental content, they do so in radically different ways, and consequently they have qualitatively different functions in mathematical cognition. Symbols represent by convention and allow mental computations to be replaced by epistemic actions. Figures and diagrams both serve as material anchors for conceptual structures. However, figures do so by having a direct likeness to the objects they represent, whereas diagrams have a metaphorical likeness. Thus, I claim that diagrams can be seen as material anchors for conceptual mappings. This classification of diagrams is of theoretical importance as it sheds light on the functional role played by conceptual mappings in the production of new mathematical knowledge.

1 Introduction

After the formalistic ban on figures, a renewed interest in the visual representation used in mathematics has grown during the last few decades (see e.g. [11–13, 24, 27, 28, 31]). It is clear that modern mathematics relies heavily on the use of several different types of representations. Using a rough classification, modern mathematicians use: written words, symbols, figures and diagrams. But why do mathematicians use different representational forms and not only, say, symbols or written words? In this paper I will try to answer this question by analyzing the cognitive function of the different representational forms used in mathematics. Especially, I will focus on the somewhat mysterious category of diagrams and

M. W. Johansen (✉)
Københavns Universitet, Copenhagen, Denmark
e-mail: mwj@ind.ku.dk

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_6, © Springer-Verlag Berlin Heidelberg 2014

explain why diagrams can be seen as material anchors for conceptual metaphors and blends. However, in order to identify what is special about diagrams, we will have to analyze the other representational forms as well. Thus, I will begin by explaining the cognitive significance of symbols and figures, and conclude with the analysis of diagrams.

2 The General Function of External Representations

In modern cognitive science there is a growing understanding of the fact that human cognition cannot be understood only by looking at the processes going on inside the human brain. As it is, our cognitive life seems to involve the external environment in several important ways.

From this point of view, external representations have several different cognitive functions. They obviously serve communicative purposes and reduce demands on internal memory. Furthermore, as noticed by Andy Clark [5], when thoughts are represented in an external media they are in non-trivial ways turned into objects. This objectification of thoughts allows us—and others—to inspect and criticize the thoughts. We can carefully scrutinize each step of a complicated argument and even form meta-thoughts about our own thinking. It seems that the use of external representations of mental content is a prerequisite for the formation of high-level cognitive processes. This function is clearly relevant in the case of mathematics, both in regard to the everyday practice of working mathematicians and in regard to the formation of meta-mathematics; in fact, meta-mathematics is exactly the kind of high-level cognitive processes that depends crucially on our ability to represent mathematical thoughts in an external media.

The last general function of external representations I will go through here is their ability to serve as material anchors for complex thoughts. As noticed by Edwin Hutchins [15], our ability to perform reasoning involving complex conceptual structures depends on our ability to represent such structures in an appropriate way. When we reason with a complex structure, we manipulate parts of the structure while the rest is kept stable. Unfortunately, our ability to do this mentally is limited. Consequently, human reasoning can be facilitated by the use of external representations that allow us to anchor some of the elements of the conceptual structure in physical representations that are globally stable, but locally manipulable. The combination of stability and manipulability of some external representations allows us to focus on the part of the structure manipulated on, while the rest of the structure is kept stable by the external media. Consequently, we might be able to work on more complex conceptual structures and perform more complex manipulations if we use an appropriate external representation. As we shall see, this anchoring property plays an important part in the function of both figures and diagrams.

It should be noted that external representations have several other important cognitive functions, see e.g. [20] for an overview.

3 Symbols and Words

3.1 Symbols as Semantic and Syntactic Objects

It is a well-known fact that mathematical symbols can be treated as both semantic and syntactic objects (see e.g. [7, 8]). As semantic objects, symbols carry mathematical content or meaning. With a few rare exceptions modern mathematical symbols carry meaning only by convention; the symbols are abstract and have no likeness with their referents. The symbol “·” for instance, does not resemble the arithmetic operation of multiplication, and the Hindu-Arabic numeral “8” does not have any likeness with the quantity eight (such as for instance the Egyptian symbol “IIIIIIII”, where the token “I” is repeated eight times).

As syntactic objects, symbols are objects of formal transformations. When symbols are treated as syntactic objects, the meaning of the symbols is suspended, and the problem at hand is solved by manipulating the symbols following purely formal rules. As an example, consider how the points of intersection between a circle and a straight line are found using analytic geometry. If we are to find the intersection points between a straight line l with slope $-1/2$ going through (0;10) and a circle C with center (4;3) and radius 5, we will at first have to find the equations of the two objects ($y = -1/2x + 10$ and $(x-4)^2 + (y-3)^2 = 5^2$, respectively). Then we must substitute the y in C’s equation with the expression $-1/2x + 10$, simplify and solve the resulting quadratic equation. During the solution process, we are not interested in—and we do not use—the geometric interpretation of the symbols. We are only concerned with the actual symbols on the paper before us. There is no reference to the meaning or content of the symbols, only to the symbolic forms and the transformations we make on them; we talk about “substituting” one symbolic form with another and “simplifying“ other expressions. The meaning of the symbols is only restored when the solution is found and given a geometric interpretation in the form of intersection points.

Arguably, this is a simple example taken from high-school mathematics, but the same dialectic between the use of symbols as semantic and syntactic objects can be found in more advanced mathematical texts as well (for analysis of more elaborate examples see [18, p. 139]).

In this case the symbols are in cognitive terms used as *cognitive artifacts* allowing computations to be performed as *epistemic actions* (cf. [7, 18, 21]). For those not familiar with distributed cognition, this might call for some explanation. An epistemic action is an actions taken in order to get information or solve computational tasks, and not in order to reach a pragmatic goal. The use of epistemic actions is a well-known cognitive strategy. When, for instance, we solve a jigsaw puzzle, we rotate and manipulate the external, physical pieces of the puzzle in order to *see* where they fit. In other words, we solve a computational problem by performing motor actions and perceiving the results of those actions. In theory we could solve the problem mentally by make internal models of the pieces and think out a solution, but we rarely do that. The reason is simple.

As David Kirsh has put it: “Cognitive processes flow to wherever it is cheaper to perform them [20, p. 442]”, and for humans it is cognitively cheaper, faster and more reliable to solve jigsaw puzzles by performing epistemic actions on the external puzzle pieces rather than thinking out a solution.[1]

In other cases, however, our physical environment does not support the externalization of a problem. For this reason we sometimes produce special artifacts—cognitive artifacts—that allow us to externalize the problem and solve it by performing epistemic actions. When treated as syntactic objects, mathematical symbols are exactly such cognitive artifacts: They allow us to substitute mental computations with epistemic actions, and that is the cognitively cheapest way of solving some mathematical problems.

3.2 Symbols as Physical Objects

As mentioned above, it is well-known that symbols can be used as semantic and syntactic objects. I would, however, like to point out that symbols can also play a third and qualitatively different role in mathematical cognition, and that is the role as plain physical objects. This use of symbols is often manifested in pen-and-paper calculations. When for instance we multiply two numbers using pen and paper, the results of the sub-calculations are carefully arranged in columns and rows, and we use the previously written results as visual cues as to where we should write the next sub-result. In other words, the physical layout of the symbols is used as a way to guide the (epistemic) actions we perform on them (see [18, p. 125] for more elaborate examples).

The use of symbols as objects is also clear in the case of matrix multiplication. Here, the usual arrangement of the elements of matrices in columns and rows is a considerable help when we have to locate the elements we are about to operate on in a particular step in the process (cf. [11, p. 242]) . Notice, that the algebraic structure of matrix multiplication is completely independent of the usual physical layout of the symbols; the product of an $m \times n$ matrix $A = [a_{ij}]$ with an $n \times p$ matrix $B = [b_{jk}]$ can simply be defined as the $m \times p$ matrix, whose *ik*-entry is the sum:

$$\sum_{j=1}^{n} a_{ij} b_{jk}$$

(see e.g. [29, p. 178]). So in theory, it would be possible to perform matrix multiplication on two unsorted lists A and B of indexed elements. In that case, it

[1] It should be noted that there is an interesting parallelism between the concept of *epistemic actions* developed in [21], and the concept of *manipulative abduction* developed by Lorenzo Magnani (e.g. [25, 26]). Magnani’s concept is however developed in a slightly different theoretical framework, and it would take us too far astray to explore the parallelism further.

would however pose a considerable task to find the right elements to operate on. By arranging the elements of the matrices in columns and rows in the usual way, the cost of this task is markedly reduced: The sum given above is simply the dot product of the ith row of A and the kth column of B, and if you know that, it is easy to find the elements you need. Thus, the multiplication process is clearly guided by the physical layout of the matrices.

It has also been suggested that the actual physical, or rather: *typographical* layout of symbolic representation of mathematical content has in some cases inspired new theorems and theoretical developments. Leibniz' derivation of the general product formula for differentiation is a case in point. Using the standard symbolism, the formula can be stated as:

$$d^n\overline{(xy)} = d^n x d^0 y + \frac{n}{1} d^{n-1} x d^1 y + \frac{n(n-1)}{1 \cdot 2} d^{n-2} x d^2 y \ etc. \tag{1}$$

It has been suggested that Leibniz' derived the formula by making a few, inspired substitutions in Newton's binominal formula:

$$^n\overline{(x+y)} = x^n y^0 + \frac{n}{1} x^{n-1} y^1 + \frac{n(n-1)}{1 \cdot 2} x^{n-2} y^2 \ etc. \tag{2}$$

(see [18, 24, p. 155] for further elaboration and more examples).

3.3 Words as Abstract Symbols

Finally, we might compare the use of abstract mathematical symbols with the use of written words. It should be noted that words are also abstract symbols (at least in alphabetic systems). In general, the physical appearance of a word has no likeness with the object, the word is supposed to represent; The word-picture "point", say, does not look like a point, and the word-picture "eight" does not have any more likeness with eight units than the abstract number-symbol "8".

Furthermore, words can carry mathematical content just as well as mathematical symbols. Of course, in general symbols allow a much shorter and more compact representation of a given content, but that is, in my view, only a superficial difference between the two representational forms. The important difference between written words and symbols is the fact that mathematical symbols, besides their role as bearers of content, can also be treated as syntactic and physical objects. With a few rare exceptions (such as avant-garde poetry), written words are never used as more than semantic objects; they cannot be used for purely syntactic transformations or as purely physical objects. For this reason, there are qualitative differences between written words and written mathematical symbols. We can simply do more with symbols than we can do with written words (cf. [18, p. 136]) .

4 Figures

4.1 Figures as Anchors of Conceptual Structures

Let me start with an example from classical geometry. In Heath [14, p. 197] Euclid's proof of Pythagoras' theorem (Euclid I.47) is accompanied by a figure similar to Fig. 1. The figure represents a particular construction that is used in the proof. Interestingly, the figure is not the only representation of the construction. During the course of the proof we are also give a full verbal description of the construction (see caption of Fig. 1). So why do we need the figure?

Most external representations can be said to anchor conceptual content, but the thing is that figures anchor content in a qualitatively different way than rhetoric and symbolic representations. Figures are holistic objects that present themselves as immediately meaningful to us. Thus, figures do not only provide a material anchor for the conceptual structure at hand; they provide an anchor that grounds our understanding of the conceptual structure in every-day sensory-motor experience of the physical world. In order to understand and give meaning to the content of the proof of Euclid I.47 we simply need a figure such as Fig. 1. We could of course imagine the figure or construct it in our mind's eye, but by doing so the limits of our short-term memory would pose limits to the complexity of the proofs we were able to understand. To use the example at hand, most people would, I believe, find it hard to keep track of the twelve individual points involved in the construction used in Euclid I.47, if they were only given a verbal description. By drawing a figure using a (semi)-stable medium we reduce the demands on short-term memory and are thus able to increase the complexity of the conceptual

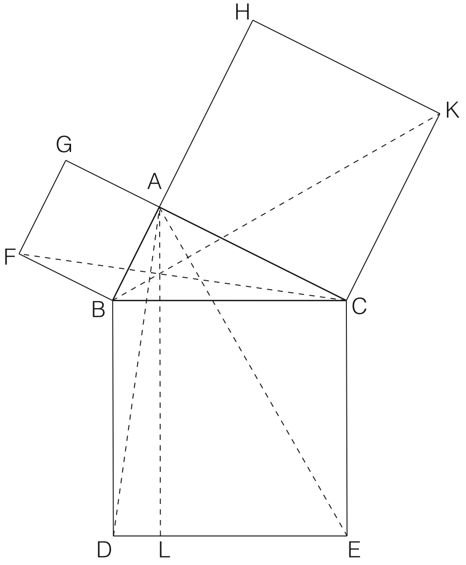


Fig. 1 Figure form the proof of Pythagoras' theorem (Euclid I.47) (redrawn from [14, p. 197]). In the proof, the following verbal description of the construction is given: "Let *ABC* be a right-angled triangle having the angle *BAC* right. [... L]et there be described on *BC* the square *BDEC*, and on *BA*, *AC* the squares *GB*, *HC*; through *A* let *AL* be drawn parallel to either *BD* or *CE*, and let *AD*, *FC* be joined. From this construction the proof proceeds ([14, p. 197])"

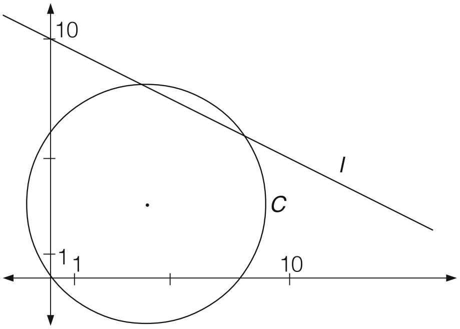
structures we are able to handle. That is one of the reasons why we use figures as a material anchor for complex conceptual structure, such as the structure involved in Euclid I.47.

4.2 Knowledge Deduced from Figures

Apart from making it easier for us to grasp a given mathematical construction, figures can also in some cases be used to deduce information about the mathematical objects represented in the figure. As a first example, we can return to the circle C and line l discussed in Sect. 3 above. There, I gave an outline of how the intersection points between C and l could be determined by analytic means. Another—and perhaps more direct—strategy would be to draw the two objects and simply read the intersection points off from the Figure (see Fig. 2). As we do not need to use information about the intersection points in order to construct such a figure, the figure clearly allows us to deduce genuinely new information about the objects it represents.

As another and slightly different example, we can look at the very first proof of *The Elements*. Here, Euclid shows how to construct an equilateral triangle on a given line segment AB (see Fig. 3). In order to achieve this result Euclid constructs two circles, one with center A and radius AB and another with center B and radius AB. Euclid then proves that the wanted equilateral triangle can be constructed by using the intersection point C of the two circles as the third vertices in the triangle. From a philosophical point of view, the interesting thing about this proof is the fact that Euclid does not *prove* that the two circles have an intersection point (we now know that the existence of this intersection point cannot be proven from Euclid's axioms). However, on the figure we can clearly *see* that the two circles have an intersection point, as any circles constructed in this way must have. So once more, we can deduce more information from the figure than we used in its construction, and apparently this was precisely what Euclid did. So here, information deduced from a figure plays an indispensable part of a mathematical proof.

Fig. 2 Figure representing intersecting line and circle



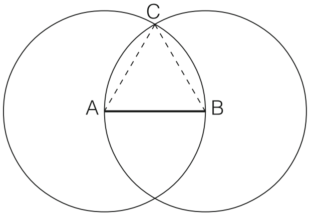


Fig. 3 Euclid I.1: How to construct an equilateral triangle on a given line segment *AB*

The use of figures in mathematical reasoning is hotly debated, and there are several things to discuss in connection to the three examples given above. At first, it should be noted that the use of figures as an aid to grasp mathematical content (such as the figure accompanying Euclid I.47) is largely recognized and supported, even by formalistically inclined mathematicians. Moritz Pasch for instance readily admits that figures "can substantially facilitate the grasp of the relations stated in a theorem and the constructions used in a proof [30, p. 43], my translation." So, even by the standards of Pasch it is legitimate to use a figure to anchor the conceptual content of a mathematical construction.

The controversy only begins when we move to Figs. 2 and 3 above. Here, the figures are used not only to illustrate, but also to infer new mathematical knowledge. The question is whether we can trust this knowledge. What is the epistemic status of knowledge deduced from a figure?

Let us begin by discussing the quality of the knowledge deduced from Fig. 2. If we compare the analytic and the pictorial method of determining intersection points, it is clear that the information deduced from the figure is not as precise as the information obtained by analytic means. Consequently, this use of figures is mainly considered a heuristic tool, and any information deduced from a figure should be tested by more reliable (i.e. analytic) means. So for instance, in the Danish high-school system a figure such as Fig. 2 is considered a valid method of finding intersection points between two curves, but the solutions read off from the figure should be tested (by means of the analytic expressions of the curves in question) if the solution is to count as a satisfying answer to the problem.

If we move to Fig. 3, the negative evaluation of the knowledge deduced from the figure is even stronger. Here, the knowledge deduced from the figure is used as an essential step in a mathematical proof, and that is—by several parties—considered an illegitimate use of pictorial knowledge. Pasch for instance continues the quote given above by stating that:

> If you are not afraid to spend some time and effort, you can always omit the figure in the proof of any theorem, indeed, a theorem is only really proved if the proof is completely independent of the figure. [...] A theorem cannot be justified by figure considerations, only by a proof; any inference that appears in the proof must have its counterpart in the figure, but the theorem can only be justified by reference to a specific previously shown theorem (or definition), and not by reference to the figure ([30, p. 43], my translation).

So Pasch would not accept Euclid's proof of Euclid I.1 as legitimate because a vital step in the proof depends on knowledge deduced from a figure. Pasch was not alone in this assessment of figures. It is well-known that David Hilbert shared

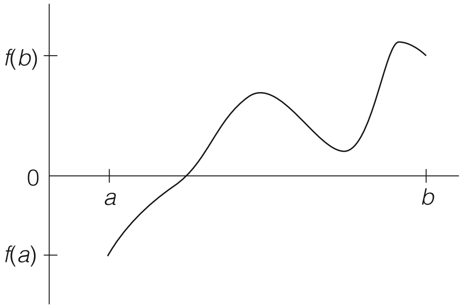


Fig. 4 Picture proof of the intermediate zero theorem. The theorem states that if a continuous function $f(x)$ defined on the interval $[a; b]$ takes both positive and negative values on the interval, then there exists a $c \in [a; b]$ such that $f(c) = 0$

Pasch's viewpoint on this matter (see e.g. [27]), and in the formalist movement a proof is in general considered "a syntactic object consisting only of sentences arranged in a finite and inspectable array [34, p. 304]". It goes without saying that figures do not have a place in such an array.

In recent years this negative evaluation of knowledge deduced from figures has been challenged by amongst others by Marcus Giaquinto [11, 12], Brown [4] and Davis [6]. Thus, Giaquinto accepts Euclid's use of a figure in the proof of Euclid I.1 on the ground that: 1) the inferences drawn from the figure does not depend on exact properties of the figure and 2) the subject matter of the proof is a homogenous class of mathematical objects (circles) that have a close relationship to a perceptual concept (perceptual circles, as the ones seen in the figure). Brown and Davis are both even more liberal in their use of figures. Brown [4, p. 25] considers Fig. 4 an adequate proof of the intermediate zero theorem (something explicitly rejected by Giaquinto), and Davis considers Fig. 5 to constitute a valid proof of the theorem that you cannot cover a circle with a finite number of smaller, non-overlapping circles [6, p. 338].

4.3 Figures and Objects

There are a number of well-known problems connected to the use of figures in mathematical deductions. As noted above, figures might not have the necessary precision, and consequently proofs based on figures can be misleading (as Rouse

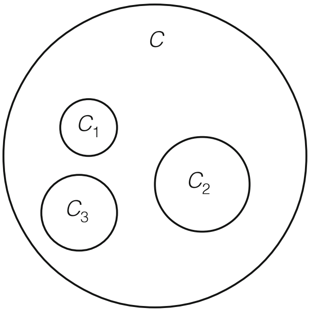


Fig. 5 Picture proof that you cannot cover a circle with a finite number of smaller, non-overlapping circles (redrawn from [6, p. 338])

Ball's famous proof that all triangles are isosceles potently illustrates [2, p. 38]). Furthermore, figures are in some cases over-specified (i.e. you cannot draw a general triangle, only a specific one) and in others they lack generality (see [11, p. 137] for more). However, none of these problems can, in my view, justify a complete ban on the use of knowledge deduced from figures. They merely impose limitations that should be observed (see also [3]).

To my mind, the main problem concerning the use of figures is connected to another and cognitively more interesting question: Why can we apparently use figures to deduce knowledge about mathematical objects *at all*?

The simple and straightforward answer is that figures somehow resembles the mathematical objects they represent; the circles drawn in Fig. 3 simply have a likeness with mathematical circles. Unfortunately, this intuitive idea is faced with several problems. Firstly, it seem to presuppose the existence of mathematical objects, or in other words Platonism, and secondly even if this presupposition is granted, it is not clear what it would mean for a physical drawing to resemble a platonic object. As a way to avoid these problems, I suggest that we see things slightly differently. Some mathematical entities such as circles and triangles are not pre-existing objects, but rather concepts created by us. They are not created at random, but are rather abstractions from and idealizations of classes of perceptual objects and shapes (see [18, p. 163]. See also [12] for a cognitively realistic account of how such an abstraction process might in fact be carried out).

Seen in this light, the longer and more correct way of explaining the relationship between a figure and the mathematical objects it represents, is the following: The figure has a direct likeness with the members of the general class of perceptual objects that provide the abstraction basis for the mathematical objects, the figure represents. Or better still: One could say that some mathematical objects such as circles and triangles are attempts to model certain aspects of physical reality, and that the shapes we see on the figures above have a direct likeness with the physical objects, the mathematical concepts are used to model. Of course not all mathematical objects have such a direct connection to sense perception, but then again: not all mathematical objects are naturally represented using figures.

If we see the relationship between perceptual figures and mathematical objects as suggested above, the real epistemological problem connected to the use of figures becomes clear. If I use information deduced from a perceptual figure in order to prove a theorem, I have proved the theorem for the wrong kind of objects. I have proven that the theorem holds good for perceptual objects, but not for the corresponding mathematical ones. Although the mathematical objects are supposed to model the perceptual objects, they might not do so perfectly; there might be a mistake in the model. In the cases above, we might accidentally have defined mathematical circles and functions in ways that would allow them to have holes where the corresponding perceptual figures have intersection points—a function such as $f(x) = x^2 - 2$ defined only on the rational numbers is a very potent example of such an object (as also pointed out by [12]). The function changes sign on the interval $[0;2]$, but there is no rational number $c \in [0;2]$ such that $f(c) = 0$. So the intermediate zero theorem does not hold good for this particular function. So it seems that we should be careful when we draw conclusions about mathematical objects on the basis of a perceptual figure. At least, we should make sure that the mathematical objects model the relevant properties of the figure in the right way.

This observation on the other hand does not show purely deductive proof to be epistemologically primary to proofs relying on figures. In my view, a figure can provide ample proof that a theorem holds good for a class of perceptual figures, and consequently the theorem *ought* to hold good for the mathematical objects modeling the class of figures as well. One can to a certain extend see the rigorization and axiomatization of mathematics during the 19th century as an attempt to make this come true. Thus, Hilbert's and Pasch's work on geometry was not an attempt to overthrow Euclid, but rather an attempt to make explicit all of the axioms needed in order to give rigorous proofs of all of the theorems of Euclidian geometry. Furthermore, as pointed out by Brown [4, p. 25] , something similar can be said about the proof of the intermediate zero theorem. As it is, the rigorous, formal proof accepted today presupposes the completeness of the real number system (amongst other things). However, this property of the real number system was not simply discovered or chosen at random. When the real number system was rigorously constructed during the 19th century, it was given the property of completeness exactly in order to make it possible to give deductive proofs of the intermediate zero theorem and other theorems presupposing the 'gaplessness' of real numbered function. Thus, proofs about perceptual figures do not only serve as a heuristic tool helping mathematicians to identify theorems they subsequently can give deductive proofs. Proofs about figures can serve—or at least have historically served as—a way to point out some of the properties we want our mathematical objects and deductive systems to have.

5 Diagrams

Before we begin, it should be realized that the word 'diagram' is used to describe a multitude of different external representations. This class of representations is not homogenous, in part because the terms 'diagram' and 'figure' are often treated as synonymous. As I see it, there are qualitative differences between figures diagrams, and part of my goal with the following analysis is to explain these difference in order to introduce a classification of mathematical representations that is more in line with their cognitive function.

It should also be noted that a lot of the work previously done on the use of diagrams in mathematics has focused on the logically soundness of diagram based reasoning. This line of work goes back at least to C. S. Peirce (1839–1914), and was revived in the mid 1990's, in part by researchers connected to the development of artificial intelligence (see e.g. [1, 13, 32]. See also [33] for a historical overview). The main goal of this program is to create a *diagrammatic calculus*, that is: a diagram based system of representations and formal transformation rules that allows for logically valid reasoning. Judged by its own standards, this program has been a great success. Several logically valid diagram based reasoning systems have been produced and some even implemented in computer based reasoning systems. The success however, has come at a price. Most mathematicians use diagrams as a heuristic tool, but instead of describing and understanding how diagrams fulfill this role, the program has focused on creating a new and different role for diagrams by turning them into a tool for logically valid, formal reasoning. To use the terminology of Sect. 3, the program wants to use diagrams as *syntactic objects*, similar to the way symbols are (in part) used. Although valuable in itself, this largely leaves the heuristic power of diagrams unexplained. In my view, there are qualitative differences between diagrams and symbols, just as there are between diagrams and figures, and if we want to understand the role diagrams play in human reasoning, we should acknowledge these differences and not shape diagrammatic reasoning into the paradigm case of logically valid, formal deduction.

Thus prepared we should have an example. Fig. 6 is a Venn diagram. Such diagrams are typically used to represent sets, and in this case three sets, *A*, *B* and *C*, are represented. Interestingly, each set is represented as a circle, or rather: a bounded area, and such areas do not have any apparent likeness with sets. Mathematical sets are abstract objects, and as such they do not have any inherent spatial characteristics. So what is the precise relationship between a diagram and the objects it represents? In the case of figures we solved the similar problem by pointing out that a figure has a direct likeness, not with the mathematical object it represents, but with the class of objects that forms the abstraction basis for the mathematical object. Unfortunately, this strategy does not seem to be viable in the case of diagrams. If we look to the objects that form the abstraction basis for, say, the set of real numbers or the set of continuous functions, it is not clear that they are arranged in anything like a bounded area. Even in more relaxed and non-technical examples (e.g. 'the sets of nouns' or 'the set of chairs') is it in general not possible to see a likeness with bounded areas. So it seems that diagrams such as Fig. 6 represents mathematical objects in a qualitatively different way than figures. The question is: How? Why do we see Venn diagrams as a representations of sets at all?

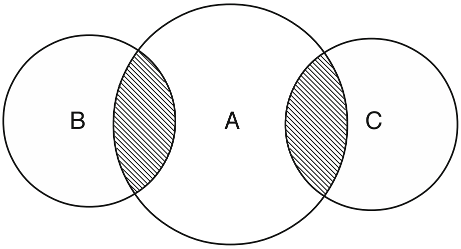


Fig. 6 A typical Venn diagram (redrawn from [16, p. 4])

In the diagrammatic calculus-program mentioned above it has been suggested that there is (or rather: ought to be) a homomorphism between a diagram and the objects it represents (e.g. [3]). The main function of this description is to explore the validity of diagrammatic reasoning, not to explore its cognitive nature. Consequently, the description does not explain why we effortlessly see certain diagrams as representations of certain mathematical objects, nor does it explain the heuristic power of diagrams. If we want to understand how diagrams represent objects, we must in my view explore the cognitive function of diagrams.

As I see it, we combine two different cognitive strategies when we use diagrams. Diagrams are external representations, and as such they can be seen as an expression of the general cognitive strategy of externalization of mental content, described above. All of the representational forms considered in this paper are expressions of this strategy. In other words, it is the second strategy that sets diagrams apart, and that strategy is *conceptual mapping*.

Conceptual mapping is a general cognitive mechanism where either one conceptual domain is mapped onto another, or a third, fictive domain is created by integrating two different conceptual domains. The first mechanism is usually referred to as 'conceptual metaphor' and the second as 'conceptual blending'. Both mechanisms are well described in the literature (see e.g. [10, 22]), and their cognitive function in mathematics has been discussed (e.g. [18, 23]), although focus has mainly been on linguistic expressions of such mappings. The main focus and contribution here will be to expand the analysis to cover non-linguistic expressions of conceptual mappings. I will do this by analyzing the role played by conceptual metaphors in our use of diagrams.

5.1 Elements of the Container Metaphor

In modern mathematics a multitude of different conceptual metaphors are in use. One of these is the Sets Are Containers-metaphor, where sets are conceptualized as containers. It is not hard to find linguistic expressions of the metaphor; most introductions to set theory will one way or the other conceptualize sets as containers. As an example, we can look at the textbook *Basic Algebra* by Nathan Jacobson. Here, sets are introduced as arbitrary collections of elements, and the basic properties of and relations between sets are described in the following way:

> If A and $B \in \mathcal{P}(S)$ (that is, A and B are subsets of S) we say that A is *contained in* B or is a *subset of* B (or that B *contains* A) and denote this as $A \subset B$ (or $B \supset A$) if every element a in A is also an element in B. [...] If A and B are subsets of S, the subset of S of elements c such that $c \in A$ and $c \in B$ is called the *intersection* of A and B. We denote this subset as $A \cap B$. If there are no elements of S contained in both A and B, that is, $A \cap B = \emptyset$, then A and B are said to be *disjoint* (or *non-overlapping*) ([16, pp. 3–4]. All emphasis form the original).

Here, sets are clearly described as containers: A set B can *contain* another set A, and both sets can *contain* elements *etc*. However, as sets are arbitrary collections of objects, they cannot literally contain anything. Thus, the description of sets as containers must be metaphorical. To be more precise, it is an example of a common conceptual metaphor, where properties from one domain—containers—are mapped onto another domain—sets: in more detail, sets are understood as containers, subsets as containers located inside a container, and the elements of a set as objects contained in a container (see [18, p. 174] for further details).

From a cognitive point of view, the main function of this conceptual mapping is to allow us to ground our understanding of sets in our experience of containers. As it is, most humans have constant experiences with containers. We use containers such as bottles and boxes an a daily basis, we move containers around and contain them in other containers (as when we put a bottle in a bag or a Tupperware in the refrigerator). We are ourselves contained in clothes and buildings and contain the food and liquid we consume. So in sum, we know a lot about containers from direct experience (cf. [19, p. 21]).

Mathematical sets on the other hand are abstract objects. We cannot experience sets directly in any way, but the conceptualizing of sets as containers allows us to map our experiences with containers onto the domain of sets. This move gives us an intuitive grasp of sets and—more importantly—allows us to recruit our knowledge about containers when we reason about sets. Conceptual metaphors are inference preserving, so if we know something to be true about containers from direct experience, we can simply activate the metaphor and map the conclusion onto the domain of sets. When sets are understood as containers, it is for instance easy to see that if a set *A* is enclosed in another set *B*, all the elements of *A* will also be elements of *B*, because we know this to be true of the content of a container *A* enclosed in another container *B*.

5.2 Diagrams as Material Anchors for Conceptual Mappings

It is now time to return to Fig. 6, were three sets were represented as bounded areas. As noted above, bounded areas do not have any direct likeness with sets. Consequently, the circles making up the bounded areas could be seen as abstract or conventional representations, similar to abstract symbols. There is however something more at play in the diagram. Circles, or rather: bounded areas constitute a special type of containers. Unlike some containers, bounded areas can overlap, but otherwise they are encompassed by the same basic logic as containers in general. Now, if we use the Sets Are Containers-metaphor to conceptualize mathematical sets as containers, we can see that the circles making up the bounded areas of Fig. 6 are more than arbitrary representations. The circles might not have a direct likeness with mathematical sets, but they do have a direct likeness with containers, and when we conceptualize sets as containers, the diagram gets an indirect or *metaphorical* likeness with mathematical sets as well.

In order to understand and use the diagram we must in other words conceptualized the mathematical objects, the diagram represents, using a particular conceptual mapping. This, in my view, marks the principal qualitative difference between diagrams and the other representational forms discussed above: In contrast to symbols, diagrams are not abstract representations, and in contrast to figures, diagrams only have an indirect likeness with the objects, they represent.

Furthermore, we should be aware that Fig. 6 is not only a representation of three sets. It is in fact a diagrammatic proof of the distributive law for set operations $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (cf. [16, p. 4]). It can easily be seen that the shaded areas of the diagram corresponds to both $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$. Thus, the identity holds good for bounded areas of the plane, and by using the Sets Are Containers-metaphor, we can map this inference onto the domain of sets to get the corresponding set-theoretical identity.

The use of a diagram to support this kind of picture proof draws attention to the double nature of diagrams. Diagrams are an expression of our use of conceptual maps, but they are also external representations, and as such shares some of the properties of figures and symbols. In particular, diagrams are, similar to figures, able to function as material anchors for conceptual structures. In a diagram, a conceptual structure established via a metaphor or conceptual blend is mapped onto an external, physical structure, whose individual elements serve as proxies for elements of the conceptual structure. Furthermore, the physical structure is globally stable, but locally manipulable. When a diagram is drawn, we can for instance add new bounded areas to it or, as it is the case in Fig. 6, shade certain areas without altering the overall structure of the diagram. By manipulating and inspecting a diagram we can, due to the underlying conceptual mapping, draw inferences about the objects represented by the diagram (or rather: about the objects as they are conceptualized under the given metaphor or blend). In the case above, the distributive law was for instance verified simply by drawing and inspecting Fig. 6. As in the case of figures, this anchoring property allows us to increase the complexity of the conceptual structures we are able to work on. So to give the full cognitive characterization, diagrams are material anchors for conceptual mappings. In this way, diagrams are—from a cognitive point of view—highly complicated cognitive tools, and they are clearly qualitatively different from both figures and symbols.

As noted above, inferences based on the inspection of a figure can for several reasons lead to false conclusions. When we turn to inferences based on diagrams, we must add a new entry to the list of possible errors, and that is: inadequacy of the metaphor. When we use a diagram to draw inferences, we reason about the mathematical objects taken under a particular metaphorical conception, but unfortunately metaphors can be misleading (in fact, some authors claim they always are (see e.g. [17]). In the case of Venn diagrams, the underlying container metaphor is clearly inadequate in several respects. Firstly, the metaphor does not capture the mathematically important difference between a set A being a subset of another set B and A being an element of B. Secondly, the metaphor might lead to false conclusions about the relative size of infinitely large sets. If for instance the natural numbers and the integers are represented in a Venn style diagram it would seem from the diagram that the set of integers is larger than the set of natural

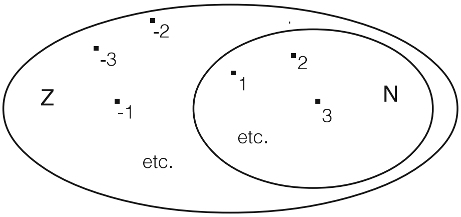


Fig. 7 The natural numbers and the integers

numbers (see Fig. 7). However, from a mathematical point of view, it isn't. This observation is of course extremely important to keep in mind when one is reasoning with diagrams.

5.3 Commutative Diagrams

So far, the analysis has only been based on one type of diagrams: Venn diagrams. Such diagrams are well-suited as examples because they are easy to understand and can be used to represent mathematically basic results. On the other hand, working mathematicians rarely use Venn diagrams, so in order to show the generality of the analysis, we should also cover a mathematically more realistic example. For this reason I will make a brief analysis of one of the diagrammatic forms most commonly used in modern mathematics: the commutative diagram (see also for more examples [18]).

In short, commutative diagrams are diagrammatic representations of maps between sets (where the exact type of maps and sets are typically specified further). Following Jacobson once more, from a strictly formal point of view a *map* consists of three sets: a domain S, a codomain T and a set α of ordered pairs (s, t), with $s \in S$, $t \in T$ and such that:

1. for any $s \in S$ there exists a pair $(s, t) \in \alpha$ for some $t \in T$.
2. if $(s, t) \in \alpha$ and $(s, q) \in \alpha$ then $t = q$ [16, p. 5].

Less formally stated, a map is a set of relations between the elements of two sets.

In modern mathematics maps are commonly represented as an arrow going from one symbol representing the domain to another symbol representing the codomain. Thus, the map f with domain A and codomain B can for instance be represented as: $A \xrightarrow{f} B$. This representation draws on several conceptual metaphors. In order to understand the diagram, the two sets must be conceptualized as locations in space, and the map must be understood as a movement along a directed path from one location to the other. As sets are not literally locations in space and maps are not literally movements, these conceptualizations are clearly metaphorical. Now, an arrow does not have a direct likeness with directed movement in space. An arrow is a (more or less) arbitrary symbol used by convention to signify movements in space. So in contrast to Fig. 6, the diagram above contains both conventional symbols and figural elements that have a direct likeness to the corresponding elements of the metaphor. Consequently, it is what I will call a *mixed anchor*.

The representation of maps as arrows between sets is especially useful when several maps between several sets are involved. So, three maps between three sets can be represented by the triangle displayed in Fig. 8.

Clearly, Fig. 8 is also a mixed anchor containing both conventional symbols and figural elements. Notice also that the symbols A, B and C are used not only as semantic objects designating the three sets, but also as purely physical objects marking the metaphorical location of the sets in the diagram.

An interesting aspect of this kind of diagrams is the fact that the conceptual metaphor embodied in the diagram is inadequate in an important respect. In the diagram, a map is represented as a movement between two sets, but from a formal point of view, a map is not a relation between two sets, but between the elements of two sets. This inadequacy has important consequences. According to Fig. 8, I can get from A to C in two ways: I can either go by the f—and then by the h-arrow, or I can go directly by taking the g-arrow. From this, it seems that the composition $h \circ f$ of f and h is equal to g. However, this might not hold good. From a mathematical point of view the composition of f and h is only equal to g if the composition, for any elements in A, will take me to the same element in C as g (i.e. $\forall x \in A : (h \circ f)(x) = g(x)$). If this is the case, the diagram is said to commute.

In order to complete the cognitive analysis, we should also notice that commutative diagrams are locally manipulable, but globally stable. We can easily add new locations (i.e. sets) or new arrows (i.e. maps) to the diagram without disturbing the overall stability of the representation.

Furthermore, commutative diagrams can be used to infer new knowledge about the objects, they represent. If we know, say, that both square $ABDE$ and square $BCEF$ in Fig. 9 commutes (i.e. $j \circ f = m \circ h$ and $k \circ g = n \circ j$), then from the diagram we can easily infer that the whole square commutes as well (intuitively, that you can go form A to F by any route you want).

Finally, commutative diagrams can also be used to anchor very complicated conceptual structures. A good example is the so-called five lemma. I will not go into the mathematical details, as they are inconsequential for our purposes (see e.g. [9]). The lemma states that for a certain types of sets (such as abelian groups), the map n is of a particular type, if a number of conditions are met (the rows must be exact and the maps k, m, p and q must be of particular types). The content of this lemma is often illustrated with a commutative diagram similar to Fig. 10.

Fig. 8 Representation of three maps between three sets

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "g"'] & B \arrow[d, "h"] \\
 & C
\end{tikzcd}$$

Fig. 9 Two commutative squares

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "h"] & B \arrow[r, "g"] \arrow[d, "j"] & C \arrow[d, "k"] \\
D \arrow[r, "m"] & E \arrow[r, "n"] & F
\end{tikzcd}$$

Fig. 10 The five lemma

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "k"] & B \arrow[r, "g"] \arrow[d, "m"] & C \arrow[r, "h"] \arrow[d, "n"] & D \arrow[r, "j"] \arrow[d, "p"] & E \arrow[d, "q"] \\
A' \arrow[r, "r"] & B' \arrow[r, "s"] & C' \arrow[r, "t"] & D' \arrow[r, "u"] & E'
\end{tikzcd}$$

Notice, that in this case no less than 23 different objects (10 sets and 13 maps) are involved. Although the usual conceptual metaphor (where sets are understood as locations in space and maps as paths between them) allows us to get a more intuitive grasp of the situation, it still poses considerable demands on short-term memory to keep track of all of the elements involved in the lemma. The conceptual structure is so complicated that we simply need an external anchor in the form of a physical diagram in order to stabilize it. In this sense the diagram above is similar to the figure accompanying Euclid I.47 (Fig. 1). They both illustrate how the use of external anchors allows us to increase the complexity of the conceptual structures we are able to work with. Only, Fig. 10 has an indirect or metaphorical likeness with the objects it represents, whereas Fig. 1 has a direct likeness.

6 Conclusion

In this paper I have categorized the different representational forms used in mathematics from a cognitive point of view. The analysis suggests a more careful use of language, especially in connection to the words ‘figure’ and ‘diagram’. However, the main aim of the categorization is not to police the use of language, but rather to draw attention to the fact that mathematical reasoning depends on a multitude of qualitatively different representational forms. From a cognitive point of view there are qualitative differences between written words, symbols, figures and diagrams. The different representational forms involve different cognitive processes and they play different roles in the reasoning process.

This appeal to recognize the differences between the various representational forms is also an appeal to recognize the complexity and diversity of mathematical reasoning. This complexity has not always been acknowledged. The formalistic movement clearly failed to recognize the complexity of mathematical reasoning by identifying mathematics with the use of a specific cognitive tool: symbols.

Although the formalistic movement has been challenged in recent decades, the exact quality of and difference between the various representational forms used in mathematics is still not well understood. From this paper it should be clear that there are several important differences between written words, symbols, figures and diagrams, and the main attraction of the cognitive perspective applied here is exactly the fact that it makes it possible for us to see and understand these differences.

References

1. Allwein, G., Barwise, J. (eds.): Logical Reasoning with Diagrams. Oxford University Press, Oxford (1996)
2. Ball, W.W.R.: Mathematical Recreations and Essays, 4th edn. Maxmillan and Co, London (1905)
3. Barwise, J., Etchemendy, J.: Heterogeneous logic. In: Glasgow, J., Narayanan, N.H., Chandrasekaran, B. (eds.) Diagrammatic Reasoning: Cognitive and Computational Perspectives, pp. 211–234. MIT Press, Cambridge (1995)
4. Brown, J.R.: Philosophy of Mathematics: An Introduction to a World of Proofs and Pictures. Routledge, London (1999)
5. Clark, A. Magic words: how language augments human computation. In: Carruthers, P., Boucher, J. (eds.) Language and Thought: Interdisciplinary Themes, PP. 162–183. Cambridge University Press, Cambridge (1998)
6. Davis, P.J.: Visual theorems. Educ. Stud. Math. **24**(4), 333–344 (1993)
7. De Cruz, H.: Mathematical symbols as epistemic actions—an extended mind perspective. Unpublished on-line working paper (2005)
8. de Cruz, H.: Innate Ideas as a Naturalistic Source of Mathematical Knowledge. Vrije Universiteit Brussel, Brussel (2007)
9. Eilenberg, S., Steenrod, N.: Foundations of Algebraic Topology. Princeton University Press, Princeton (1952)
10. Fauconnier, G., Turner, M.: The Way We Think: Conceptual Bending and the Mind's Hidden Complexities. Basic Books, New York (2003)
11. Giaquinto, M.: Visual Thinking in Mathematics:an Epistemological Study. Oxford University Press, Oxford (2007)
12. Giaquinto, M.: Crossing curves: a limit to the use of diagrams in proofs. Philosophia. Mathematica. **19**(3), 281–307 (2011)
13. Glasgow, J., Narayanan, N.H., Chandrasekaran, B.: Diagrammatic Reasoning: Cognitive and Computational Perspectives. AAAI Press, Cambridge (1995)
14. Heath, T.L.: The Thirteen Books of Euclid's Elements. Barnes & Nobel, Inc., New York (2006)
15. Hutchins, E.: Material anchors for conceptual blends. J. Pragmatics **37**(10), 1555–1577 (2005)
16. Jacobson, N.: Basic Algebra, 2nd edn. W.H. Freeman and Company, New York (1985)
17. Jensen, A.F.: Metaforens magt. Fantasiens fostre og fornuftens fødsler. Modtryk, Aarhus C (2001)
18. Johansen, M.W.: Naturalism in the Philosophy of Mathematics. Ph.D. thesis, University of Copenhagen, Faculty of Science, Copenhagen (2010)
19. Johnson, M.: The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press, Chicago (1990)
20. Kirsh, D.: Thinking with external representations. AI & Soc. **25**(4), 441–454 (2010)

21. Kirsh, D., Maglio, P.: On distinguishing epistemic from pragmatic action. Cogn. Sci. **18**(4), 513–549 (1994)
22. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (1980)
23. Lakoff, G., Núñez, R.: Where Mathematics comes from: How the Embodied Mind Brings Mathematics into Being. Basic Books, New York (2000)
24. Larvor, B.: Syntactic analogies and impossible extensions. In: Löve, B., & Müller, T. (eds.) PhiMSAMP. Philosophy of Mathematics: Sociological Aspects and Mathematical Practice. Texts in Philosophy, vol. 11, pp. 197–208. College Publications, London (2010)
25. Magnani, L.: Conjectures and manipulations: external representations in scientific reasoning. Mind Soc. **3**(1), 9–31 (2002)
26. Magnani, L.: External diagrammatization and iconic brain co-evolution. Semiotica **186**, 213–238 (2011)
27. Mancosu, P.: Visualization in logic and mathematics. In: Mancosu, P., Jørgensen, K.F., Pedersen, S.A. (eds.) Visualization, Explanation and Reasoning Styles in Mathematics, pp. 13–30. Springer, Dordrecht (2005)
28. Manders, K.: Diagram based geometric practice. In: Mancosu, P. (ed.) The Philosophy of Mathematical Practice, pp. 65–79. Oxford University Press, Oxford (2010)
29. Messer, R.: Linear Algebra: Gateway to Mathematics. HarperCollins College Publishers, New York (1994)
30. Pasch, M., ehn, M.: Vorlesungen über neuere Geometrie. Die Grundlehren der mathematischen Wissenschaften, vol. 23. Springer, Berlin (1882/1926)
31. Reviel, N.: The Shaping of Deduction in Greek Mathematics: a Study in Cognitive History. Ideas in Context. Cambridge University Press, Cambridge (1999)
32. Shin, S.J.: The Logical Status of Diagrams. Cambridge University Press, Cambridge (1994)
33. Shin, S.J., Lemon, O.: Diagrams. In: Zalta, Edward N. (ed.) The Stanford Encyclopedia of Philosophy, winter 2008 edn. http://plato.stanford.edu/archives/win2008/entries/diagrams/ (2008)
34. Tennant, N.: The withering away of formal semantics? Mind & Lang. **1**(4), 302–318 (1986)

Models and Points of View: The Analysis of the Notion of Point of View

Manuel Liz

Abstract The notion of points of view is crucial in scientific and technical disciplines when our knowledge is guided by models. There are very few analyses of the structure of points of view. However, we can identify two general approaches. One of them assumes as a paradigm the structure of propositional attitudes. Here, points of view are understood as having an internal structure similar to the one we find in propositional attitudes. The other approach is based on the notions of location and access. Here, the internal structure of points of view is not directly addressed. The features that are emphasised are related to the role that points of view are intended to have. Here, points of view are understood as ways of having access to the world, and to ourselves, from certain emplacements. The paper has three parts. In the first one, I present these two approaches and some interesting developments inside each one of them. In the second part, I examine more closely the relationships between the two approaches and defend the non-reducible, relational nature and modal character of points of view. In the third part of the paper, I draw some conclusions. The analysis of the notion of points of view is relevant by itself. But it also entails important epistemological and methodological consequences. Insofar as points of view have a non-reducible mode of existence, references to points of view in scientific and technical fields cannot be seen merely as a "second-class" kind of discourse.

This paper has been supported by Research Projects FFI2008-01205, "Points of View. A Philosophical Investigation", and FFI2011-24549, "Points of View and Temporal Structures" (Ministerio de Economía y Competitividad. Gobierno de España).

M. Liz (✉)
Universidad de La Laguna, La Laguna, Spain
e-mail: manuliz@ull.es

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_7, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

The notion of points of view, or perspectives, is important in many contexts. And it is particularly relevant in scientific and technical disciplines when our knowledge is guided by models. A model offers a certain point of view from which we get cognitive or practical advantages. Models offer peculiar points of view which help us to see and conceptualise reality and act upon it.

There are very few analyses about the structure of points of view. However, we can identify two general approaches. One of them assumes as a paradigm the structure of propositional attitudes. Points of view are understood as having an internal structure similar to the one we find in propositional attitudes. Points of view would be structurally constituted by a subject, a set of contents and a set of relations connecting the subject to those contents. This approach can have a strict conceptualist interpretation, but it can also assume the existence of non-conceptual contents. The other approach is based on the notions of location and access. The internal structure of points of view is not directly addressed. The features that are emphasised are related to the role that points of view have. Points of view are understood as ways of having access to reality from certain emplacements. A point of view would be constituted by a location offering a certain perspective.

The paper has three parts. In the first part, I present the two above mentioned approaches, and some interesting developments within each one. In the second part, I examine the relationships between these approaches, discussing some relevant implications with respect to the nature of points of view. The third part contains some conclusions. The analysis of the notion of point of view is important by itself. But it also has very important epistemological and methodological consequences. I will defend a non-reductivist conception of points of view. Points of view have an irreducible relational nature. And they also have a not less irreducible modal dimension. I will argue that points of view are not reducible to psychology, nor to information, nor to physics. Insofar as points of view have such a non-reducible mode of existence, references to them in scientific and technical fields cannot be seen as having only a heuristic or pragmatical value, and neither can models trying to incorporate those points of view be seen as constituting some kind of "second-class" discourse. If points of view are non-reducible parts of reality, then references to them have to be simply unavoidable.

2 Two Approaches to the Structure of Points of View

There is a serious lack of elaborated philosophical theories about what is the structure of points of view. However, we can identify two main approaches.

The first one assumes as a model the structure of propositional attitudes. Points of view are understood as having an internal structure similar to the one we can find in propositional attitudes. The other one is based on the notions of location and access. Here, the internal structure of points of view is not directly addressed. The features that are emphasized are related to the function that points of view have.

I will introduce some representative authors of each kind of approach. Some of these authors are not very well known. This is especially so in the case of Jon Moline and Antii Hautamäki. Both of them are, however, pioneers in the task of analyzing with philosophical depth and clarity the notion of points of view. And we can continue finding in them very important insights.

2.1 The Model of Propositional Attitudes

As I have said, propositional attitudes can offer a model for the internal structure of points of view. Points of view would be structurally constituted by a subject, a set of contents and a set of relations connecting the subject with those contents. This model can have a conceptualist reading or a non-conceptualist reading.

2.1.1 The Conceptualist Reading

According to a common analysis of propositional attitudes, they are constituted by a subject maintaining a certain psychological attitude towards a certain proposition. The proposal here is that points of view can be understood as structured sets of propositional attitudes such as “to believe”, “to desire”, “to perceive”, “to remember”, “to imagine”, “to guess”, etc., linked to certain contents.

Propositions have a conceptual structure and an inferential articulation. They are constituted by concepts and they maintain inferential relations among them. In a conceptualist reading of the model of propositional attitudes, points of view can only have contents of that kind. So the central claims of the conceptualist reading are (1) that points of view are structured sets of propositional attitudes, and (2) that the contents of a point of view are always sets of propositions conceptually structured and inferentially articulated.

The thesis that points of view are structured sets of propositional attitudes is particularly interesting when we focus on maximally large propositional fields, constituted by conceptual structures and inferential relations able to determine all possible contents. Some propositional fields are larger than other ones. Maintaining constant the inferential articulation of each field, that difference in size would depend on their respective conceptual recourses. A propositional field would be maximally large if there is no need to introduce new concepts in any situation in which some propositions have to constitute the contents of a given set of propositional attitudes. Points of view constituted by maximally large propositional fields become something very close to Carnap’s “frames”, Davidson’s “conceptual schemes”, Quine’s “manuals of translation, and Putnam’s “conceptual relativism”. Certain ways of understanding concept possession also allow us to include Kuhn's "paradigms" and Wittgenstein's "forms of live". I will not describe in detail these approaches. The important thing is that in all these approaches, the model of propositional attitudes is adopted in quite a direct way.

2.1.2 The Non-Conceptualist Reading: Christopher Peacocke

But the model of propositional attitudes can also be viewed in other very different ways. The model can be made compatible with a defence of the existence of non-conceptual contents, as something not reducible to conceptual contents.[1] Propositional attitudes can continue being a model because the analysis of the structure of points of view regarding such non-conceptual contents continues to be done in close connection with the structure of propositional attitudes.

Christopher Peacocke has recently offered an account of non-conceptual content which can be easily applied to the analysis of points of view of a non-conceptual kind. It is interesting to introduce briefly Peacocke's approach.

Peacocke ([17], Chap. 1) distinguished between representational and sensational (phenomenal, or qualitative) properties of experience, arguing that the last ones, the "what-it-is-like" to have that experience, are indispensable to characterize it. Suppose, for example, that you are seeing two trees. One of them is one hundred yards away from the other one. Your experience represents them as being the same size. However, one of them, the image of the tree that is closer to you, takes more space in your visual field than the other one. According to Peacocke [17], this is clearly a non-representational, sensational (phenomenal, or qualitative) aspect of your experience.

Peacocke [18, 19] modifies that position. Sensational properties of experience are assumed as being representational, but not in a conceptual way. Peacocke argues that experience has representational contents that cannot be individuated in a Fregean way. The Fregean criterium for the identity of concepts (senses, modes of presentation, etc.) is the following one: if two concepts C and C' are identical, then the thought that, for any x, the x that is C is the x that is C' has to be uninformative. In other words, if that thought is informative, then C and C' have to be different concepts (senses, modes of presentation, etc.). But, Peacocke says, it is a common situation to perceive, for instance, two objects as having the same length and wonder whether they really do. Here, the thought that the two objects have the same length is very informative. However, the two objects are seen as

[1] There are many ways of characterising non-conceptual content. In general, non-conceptual content is a content that cannot be described in words. More precisely, it is a content that, when it is attributed to a subject, the attribution is done in terms of concepts that the subject does not need to possess. Other characterisations of non-conceptual content involve notions like "acquaintance" [25], "continuous homogeneity" [26], "analogical content" [18], "finesse of grain" [20], the "what-it-is like" of experience [8, 14, 17], "demonstrativeness" [22], "contradictory character" [2], "know-how" [24], "agency" [15, 16], etc. With respect to all these ways of characterising the notion of non-conceptual content, see [3].

having the same length. Both of them are seen as having the same appearance. According to Peacocke, the conclusion has to be that the ways the objects are experienced, for example in visual perception, cannot be Fregean concepts (senses, modes of presentation, etc.).

In Peacocke [18, 19], there are also some other proposals about the analogical and fine-grained representational character of the non-conceptual content of experience. However, the most important development has been the notion of a "scenario content". We can find a very good presentation of that notion in Peacocke [21]. He assumes that perceptual experience represents the world as being in a certain way. Perceptual experience has a representational content. But, it is a content of a non-conceptual kind. Moreover, Peacocke claims, that non-conceptual kind of content is "the most fundamental type of representational content". The sense in which this type of content is arguably the most fundamental one is that representational properties of other sorts always presuppose in various ways the existence of this first type of content ([21], p. 105).

In order to make clear that type of representational content, Peacocke introduces the notion of "scenario content". Scenario contents are individuated "by specifying which ways of filling out the space around the perceiver are consistent with the representational contents being correct" ([21], p. 105). Scenario contents are spacial types defined by certain ways of filling out the space around the perceiver. The correctness of a scenario content is a matter of instantiation by the real world around the perceiver of the spatial type which would determine what the scenario content is representing.

Peacocke proposes the following two steps for a full specification of scenario contents:

1. The first step consists of fixing an adequate "origin" and "axes". It has to be possible for many different spatial portions of the real world to find a position in relation to that origin and axes. For instance, one such origin can be given by the centre of our chest, with three axes defined by the directions back/front, left/right and up/down defined with respect to that origin. The adequate origin and axes would be different if we consider other kinds of perceivers, for instance spherical organism, inhabitants of a fluid, whose experiences were caused by impacts of light on the surface of their bodies. Also, the adequate origin and axes are relative to each different mode of sensorial perception. The choice of an origin and axes has to capture distinctions in the phenomenology of experience itself. In other words, that choice depends on the correction conditions of the representations involved in the experience.
2. The second step in the specification of scenario contents consists of specifying the ways in which the space around the origin can be filled out. One such specification could be the following. For each point identified by a distance and direction from the origin, it is specified whether there is a surface there, and if so what orientation, texture, hue, saturation, brightness, temperature, solidity, etc., it has at that point. In the visual case, for instance, we have to include

things like the direction, intensity and character of light, the rate of change of perceptible properties, etc.

Any spacial type specified in the preceding way is called by Peacocke a "scenario", and the volume of the real world around the perceiver at the time of a certain experience is called a "scene". Scenario contents are equivalent to scenarios. And scenes are defined as what is really happening around the perceiver at a certain time. The non-conceptual content of an experience is a scenario content. And that content is correct if there is really a scene taking place in that scenario. That is, the content is correct if there is a scene at the time of the experience such that it falls under the way of filling out the space around the perceiver which constitutes the scenario.

In any case, the concepts used in making those specifications do not need to be concepts possessed by the perceiver. There is no problem using a very sophisticated conceptual apparatus to fix the scenario contents of conceptually very primitive perceivers.[2]

2.1.3 A General Reconstruction

According to the model of propositional attitudes, any point of view PoV can be seen as having the following canonical structure:

PoV = <B, R, non-CC, CC, Cp > , where

1. B is the bearer of the PoV,
2. R is a set of relations connecting B with the explicit contents of the PoV,
3. non-CC and CC are the two kinds of contents that can be explicitly included in the PoV. non-CC is a set of non-conceptual contents and CC is a set of conceptual contents, and
4. Cp is a set of possession conditions for having the PoV.

B is the entity to which the point of view can be attributed. The point of view is anchored in reality through the relations R that the bearer of the point of view B is capable of maintaining with the explicit contents non-CC and CC of the point of view. There are various relevant possibilities for B. It can be a personal subject, or

[2] I will not address further developments of Peacockes approach. Peacocke ([21], pp. 107–108) notes that in parallel with the familiar distinction in philosophy of language between the linguistic meaning of indexical expressions ("I", "here", "today", etc.) and the meaning of particular utterances of those expressions, it is possible to distinguish between an outright-assessable "positioned content" and the generic notion of scenario content. Peacocke ([21], p. 119) also introduces the notion of a "proto-propositional content" as a second layer of non-conceptual content. It is not determined by positioned content, but it does not require the possession of concepts either. Peacocke ([21], pp. 111–135) argues that the identity of conceptual contents depends upon the nature of their links with all these sorts of non-conceptual contents. That way, non-conceptual content would be capable of grounding concepts in a non-circular way.

it can be a psychological subject without the status of a person, or it can be a non-personal and non-psychological entity. Also, it can be an individual or a collectivity of individuals. The case where B is a psychological and personal individual subject like ourselves is very important. In that case, (1) the relations R become a set of psychological attitudes, (2) the non-CC of the point of view is constituted by actual or possible ordinary objects with their ordinary properties and relations, and (3) the CC of the point of view is constituted by a set of propositions conceptually structured and inferentially articulated.

Either one of the sets non-CC and CC can be empty, but not both of them. There cannot be a point of view without explicitly containing some contents, either non-conceptual or conceptual. The non-CC contained in a point of view can be construed in many ways. Peacocke has offered a very suggestive way. Other ones could be based on sense-data, or on some peculiar sorts of physical properties, etc. In any case, the result for subjects like ourselves has to be a landscape full of ordinary objects with their typical properties and relations.

The CC contained in a point of view can also be construed in many ways. For instance, propositions can be derived from concepts or, alternatively, concepts can be derived from propositions. The basic elements of CC can be either concepts or propositions. In any case, the result for subjects like ourselves has to be something having the conceptual structure and inferential articulation we can find in the contents of propositional attitudes.

The possession conditions Cp of the point of view are a constitutive part of it. And they are a very important part. Every point of view has to have some Cp. They are "internal to the point of view" in the sense that if they were not to be satisfied, then the point of view could not be maintained. In relation to those possession conditions, we can introduce the weaker notion of "attribution conditions". They allow the attribution of points of view. However, they are not constitutive to the point of view. In contrast with possession conditions, attribution conditions are "external to the point of view".

2.2 The Model of Location and Access

The second general approach about the structure of points of view focuses on the notions of location and access. Here, points of view are not internally analyzed. They are identified by their role. Points of view are understood as ways of having access to the world and to ourselves. A point of view would be constituted by a location offering a certain perspective.

I will introduce three different accounts of points of view which in one way or another follow that model. The first account can be qualified as behaviourist, the second one is a logical account, and the third one has a strong metaphysical character.

2.2.1 A Behaviourist Approach: Jon Moline

In Moline [11], we can find a very suggestive analysis of what it is to take a point of view. According to Moline, points of view are ways of viewing and considering things and events from certain locations.

Moline firmly rejects any Kantian's conception of points of view according to which points of view can be defined by certain rules which would follow from certain principles, laws, maxims, theories, etc. Rhetorically, Moline asks ([11], p. 191) "What 'principle' could be said to define the Negro point of view?". To take such a point of view does not entail any sort of principle, but to have certain attitudes linked to having a certain colour of the skin in certain social contexts. At other times, to take a point of view entails to learn some special facts, techniques or languages. This is so in cases like, for instance, that of the physicist's point of view. At other times, to take a point of view entails entertaining certain experiences of a very special sort. This is so in cases like, for instance, that of a schizophrenic point of view.

Moline [11] distinguishes two main senses of the expression "point of view". The first one is the prosaic location sense in which the expression means nothing more than a spatial location regarded as a vantage point of view of the sort a photographer might seek. The second sense is an extension of that first sense. But that extension is full of philosophical significance. A point of view would not be only a place from which one views things and events, but also the peculiar way in which those things and events can be viewed and considered from that emplacement.

Moline notes that both in the prosaic location sense and in the extended philosophical one, points of view can be either taken by only one person at a time, or they can be shared by many people. This is an important remark. Some points of view are like the top of an extremely narrow peak. Other ones, like some other peaks, are so broad that many people could be said to be looking at things from the same point of view or perspective at the same time.

Another important remark is that we can know many things about a point of view without adopting it. Furthermore, there are again two different senses in which a subject can "take", or "adopt", a point of view. There is a sense in which to take a point of view implies a certain peculiar sort of overt behaviour. And there is another sense in which the adoption of the point of view only implies a certain sort of thought. For instance, a detective in a large department store takes a detective's point of view in the first sense and takes the point of view of a shoplifter only in the second sense. He does not take the point of view of the shoplifter in any overt action, by stealing from the store. He only takes it in thought (or so we suppose), as part of a strategy of anticipation.

Moline identifies three principal sorts of claims made by using the expression "point of view" in the philosophical sense. The first sort is constituted by, what he calls, comprehension claims. They are made in explanatory contexts. One example can be the following: "If you were to try to understand her point of view, you would not find her decisions so incomprehensible". The second sort is constituted by claims about the relevance or irrelevance of certain considerations from some point of view. One example can be the following: "Psychological matters are irrelevant to logic". The third sort is constituted by size claims using qualifiers like "narrow", "restricted", "larger", "broader", "wider", and so on. One example can be the following: "He takes a narrow, economic point of view on all political issues".

In his analyses, Moline also notes that the use of the expression "point of view" is restricted by considerations of personality and role. Whereas we speak freely of things like "my personal point of view", "a paranoid point of view", "a physicist's point of view", "a bureaucratic point of view", or "a parental point of view", it is very odd to speak of things like "a coffee-break's point of view", "a cello's point of view" or "the point of view of the square root of 2". There are other cases having an intermediate character like "a dog's point of view", "a computer's point of view" or, with more oddity, "a clam's point of view". Sometimes, the appropriateness of using the expression "point of view" depends on personality (my personality, a paranoid personality, the quasi-personal character of dogs, the metaphorical personality of computers and clams). At other times, it depends on the existence of certain roles (the ones exemplified by physicists, bureaucrats, or parents). According to Moline, these connections between attributions of points of view and considerations of personality and role are very important. Both personality and role suggest what is given to a certain character to say or to do, even what it is appropriate for that character to feel, or accomplish, or assume, etc. Theatre provides a good model for that meaning.

Moline ([11], p. 194) argues for a replacement of the original question "What is it to take a point of view?" with the question "What does one taking a point of view do?". The crucial point is that the new question is not a question about any mysterious relation described precisely as taking a point of view, but a direct question about a certain expected behavior. We expect one who takes or adopts a point of view to display a set of behavioral tendencies such as the following ones:

(a) A tendency to have and pursue certain specifiable interests and aims.
(b) A tendency to use only certain criteria to evaluate actions as conducive to achieving those aims.
(c) A tendency to regard other interests and criteria as largely irrelevant.
(d) A tendency to make certain factual assumptions, but not other ones.
(e) A tendency to agree with the interests, aims, evaluative criteria and relevance judgments of others taking the same point of view.

Moline's replacement of the original question is analogous to the behavioristic change proposed some years earlier by Türing with his "simulation game". The atmosphere of those years gives support to that analogy. To take a point of view has to consist of being capable of behaving in certain peculiar ways, and sharing that point of view also has to consist of behaving in those peculiar ways.

Moline does not analyse the contents of points of view. In particular, there is no distinction between points of view with an explicit conceptual content and points of view with an explicit content which is not conceptual. Nor is there any distinction between implicit and explicit contents. Moline's approach is very behaviourist. He emphasises a notion of point of view which is largely independent of mental states. Because of that, Moline's behaviourist approach makes it impossible to distinguish between genuinely taking a point of view and merely simulating taking that point of view.

The distinction between conceptual and non conceptual points of view is very important. So is the distinction between implicit and explicit contents, and the distinction between adopting a point of view and simulating such adoption. However, there is not much room for these distinctions in Moline's behaviourist approach. Nevertheless, Moline's account is very suggestive in many other respects. The various senses in which points of view can overlap because there are some overlaps in tendencies a-e , for instance, suggest ways to overcome situations of supposed incommensurability, in particular situations of supposed relativism, without depending on any set of shared contents.

2.2.2 A Logical Approach: Antii Hautamäki

Antii Hautamäki, a Finnish logician, offered, some years ago, a very elaborated work about the notion of viewpoints. To our knowledge, it is the only research in logic dealing directly and in depth with the notion of points of view. See mainly Hautamäki [6, 7].

In his approach, "viewpoint" means a way to conceptualise the world. The main idea is that discussions about scientific change, conceptual or linguistic schemes, or frameworks, theoretical perspectives, etc., invite the creation of logics in which truth values depend not only on the world considered but also on ways of conceptualising the world. According to Hautamäki ([6], p. 187), "the transition from ordinary logic to the logic of viewpoints is a logical counterpart to the philosophical transition from [in Putnam's terms] 'metaphysical realism' to 'internal realism' ".

Hautamäki presents a propositional logic of viewpoints leaving the inner structure of points of view unspecified. Points of view are taken as propositional modal operators. The formal language consists of the usual modal operators, L (for necessity) and M (for possibility), together with two new operators: A (interpreted as "from all viewpoints p", or "absolutely p") and R (interpreted as "from some viewpoint p", or "relatively p").

The semantics for Hautamäki's logic uses Kripke models enriched with some new elements standing for viewpoints and with a new relation S defined between pairs of worlds and viewpoints. In that sense, a model is a structure $<W, I, R, S, V>$, where W is a set of possible worlds $\{w, w', \ldots\}$, I is a set of viewpoints $\{i, i', \ldots\}$, R and S are relations in WxI, and V is a function of evaluation from the set of well formed formulas F and elements of WxI to truth values $\{1, 0\}$, that is V goes from FxWxI to truth values $\{1, 0\}$. The expression Rp is defined as $\neg A \neg$ p, and the truth condition for Ap is defined as follows:

$V(Ap, w, i) = 1$ iff $V(p, w, i') = 1$, for all i' such that $\ll w, i >, < w, i' \gg$ belongs to S. This parallels the definition of Mp as $\neg L \neg p$, and the truth condition for Lp as $V(Lp, w, i) = 1$ iff $V(p, w', i) = 1$, for all w' such that $\ll w, i >, < w', i \gg$ belongs to R. What is relevant for modal operators L and M is how truth values can change in relation to changes in the possible worlds considered, and what is relevant for modal operators A and R is how truth values can change in relation to changes of perspective.

Hautamäki [6] construes various possible axiomatisations of the logic of viewpoints and he proves them to be complete. Let consider the set of modal systems $\{K, T, B, S4, S5\}$. Let x and y be any two elements of that set. Hautamäki presents logical systems of viewpoints following the structure (x, y), where the system is an x-system with respect to operators L and M in the axioms of x, and a y-system with respect to operators A and R standing for L and M in the axioms of y. According to that strategy, there are exactly 25 different systems for the logic of viewpoints.[3]

The relation S connecting pairs $< world, viewpoint >$ is crucial. How to interpret it? Hautamäki proposes understanding it as formalising the intuitive idea of "alternativeness". Two points of view would be alternative when the result of seeing the world through one of them is the same as the result of seeing the world through the other one. There are various ways to make such interpretation, some of them being more demanding than others. According to Hautamäki, the weakest adequate way consists in saying that $\ll w, i >, < w, i' \gg$ belongs to S iff $V(p, w, i) = V(p, w, i')$ for some p. With respect to some proposition p and some word w, a viewpoint i can be "alternative" to another viewpoint i'.

So defined, the relation S is reflexive and symmetric. Hence, the weakest adequate system in the set of those 25 different logical systems for viewpoints is (T, B), a logical system as T with respect to operators L and M, and as B with respect to operators A and R.

What kind of modality is introduced by the notion of points of view? Let modality be any sequence of the operators $\neg, L, M, A$ and R, including the empty sequence. Let modalities m and m' be equivalents, or reducible, in a system iff for every p, the co-implication of mp and $m'p$ is a theorem in that system. In all the logical systems for viewpoints considered by Hautamäki, the eight modalities $LA, AL, LR, RL, MA, AM, MR$ and RM are in fact distinct, or non-reducible, in the sense that equivalence fails with respect to any two of them. This entails a very important result. If we consider the possibility of combining and reiterating the above introduced modalities, then there are infinitely many distinct modalities involving points of view.

Points of view are taken by Hautamäki as propositional operators able to give place to new truth values. As I have said, this makes explicit the relativisation of

[3] Hautamäki ([6], p. 188) notes that there are close relations between his logics of viewpoints and both temporal logic and two-dimensional modal logic. He mentions, respectively, Needham and Segerberg. It would be very interesting to explore these relationships in more detail.

our knowledge claims to a point of view, or perspective. This is the core of Hautamäki's approach.

2.2.3 A Metaphysical Approach: Adrian Moore

A much more recent analysis of the notion of point of view connected with the model of location and access, this time of a metaphysical sort, can be found in Adrian Moore [12, 13].

Moore ([13], p. 6) defines a point of view as follows: "By a point of view I shall mean a location in the broadest possible sense. Hence points of view include points in space, points in time, frames of reference, historical and cultural contexts, different roles in personal relationships, points of involvement of other kinds, and the sensory apparatuses of different species".

Moore distinguishes points of view from "outlooks". It is a very important distinction in his approach. An outlook is any way of representing the world, any way of seeing it or thinking about it. When points of view are involved, our representations are dependent on a perspective. However, Moore claims, perhaps to represent the world in accord with an outlook does not entail representing it from a point of view.

Moore claims that in fact it is possible to represent the world, and our position in it, from no point of view. This is to have an "absolute representation", a representation offering an absolute outlook. Even though points of view are always perspectival, there can be representations from no point of view. Some outlooks are of that kind. They are non-perspectival outlooks.

Representations fully detached from any point of view are absolute representations. They are representations independent of any perspective. Are they really possible? The idea of an absolute and complete conception of reality comes from Williams [27]. It has been criticised by many authors (for instance, [14, 23]). But Moore defends the idea of an absolute conception as quite a direct consequence of the very notion of representing reality as "representing what there is there anyway". If there is a substantial or stable way, even minimal, in which reality is in itself, then a complete true representation having that content would be an absolute representation of reality. Different parts of that representation would be partially true, but absolute representations of reality that can be integrated through a "simple addition".

Moore's ([13], Chap. 4) argument for the possibility of absolute representations (what he calls "The Basic Argument") can be presented in the following simple way: The content of any true representation is that things are in a certain way where things are that way. But this would come to nothing if it were not possible to represent what that way is. And to represent that requires it being represented it in an absolute way. It requires representing in a sense of representing that is not dependent on any point of view.

The notion of being integrated by "simple addition" is crucial in Moore's approach. An account revealing how two incompatible representations r1 and r2 are made true by reality has to refer explicitly to the points of view involved. At least one of those representations cannot be integrated by "simple addition". At least one of them has to be integrated without adopting the relevant point of view involved. That is, at least one of them cannot be endorsed except in an indirect way. Let us consider an assertion of the sentence "It is raining" and, with respect to the same place but a time later, an assertion of the sentence "It is not raining". To explain how both representations can be true, and how they can be integrated into a single representation, entails making explicit, at least for one of them, the time at which the assertion is made.

According to Moore, the distinction absolute/perspectival is very different from the distinction objective/subjective. To think that a melody is exquisite can be an example of a representation subjective and perspectival. To say that there was a total eclipse of the Sun here yesterday can be an example of a representation objective but perspectival. In contrast, that $e = mc^2$ can be an example of a representation objective and absolute. The aim of science is to construe representations not only objective, but absolute.

Moore's approach contains many more elements, for instance a very disputable claim about the "ineffable knowledge" we can have about the absolute truth of a transcendental idealism maintaining the perspectival dependence of all our beliefs.[4] I cannot discuss here that claim. I will finish saying only something about the notion of an absolute conception of reality. Many of the ideas linked to that notion can maintain their attraction and force redefining the notion of an absolute conception of reality in terms of "independence of any particular point of view". Under this interpretation, to have an absolute conception of the reality would not be to represent it from no point of view, but to represent it with independence of any particular point of view. The key word in the notion of an absolute conception would be "perspectival invariance". Representations showing perspectival invariance would have many of the characteristic features of absolute representations. Furthermore, they can be trivially integrated by "simply addition". Perhaps, what science tries to obtain is perspectival invariance, and not representations "from no point of view".

2.2.4 A General Reconstruction

According to the model of location and access, a point of view is constituted by a peculiar location enabling the epistemic access to some facts.

Making an intuitive use of the notion of a set of possible worlds W, we can understand a point of view as a certain relation

$$We - Wp = \{ <we, wp> , <we, w'p> , <we, w''p> , \ldots \},$$

[4] A critical review of Moores theses, mainly about the supposed ineffable truth of transcendental idealism, can be found in [5].

defined in WxW and such that *we* is an emplacement-world from which we can have epistemic access to some facts concerning some worlds-in-perspective $wp, w'p, w''p$, etc.

Each point of view determines one such relation. From a certain emplacement-world, we, it would be possible to have epistemic access to some facts of some worlds-in-perspective $wp, w'p, w''p$, etc.

The approach is interesting. However, two things have to be noted. Both of them have to do with the notion of possible world. The first one is that we need to define the distinction between conceptual and non-conceptual explicit contents in terms of possible worlds. The second one is that we need to define all kinds of contextual dependence, both with respect to conceptual and non-conceptual content, also in terms of possible worlds. These two things constitute quite hard problems for any attempt of reconstructing the model of location and access in the direct way I have introduced. From this model, it is very difficult to distinguish between conceptual and non-conceptual epistemic access. Furthermore, because that model does not analyse the internal structure of points of view, we do not have anything like the relations R in order to give an account of the context dependence of many of the contents of our points of view.

Now, let us consider again the above relation,

$$We - Wp = \{ <we, wp> , <we, w'p> , <we, w''p> , \ldots \}.$$

It is very useful in order to analyse that relation between pairs of possible worlds constituted in such a way that to be located in one of them makes possible an epistemic access to facts of the other ones. However, if we do not introduce any other restriction, that relation has none of the usual formal properties that we like to have in a relation among possible worlds. It is arguable that We-Wp is a relation (1) that is not in general reflective, (2) that would only be clearly symmetrical in those cases in which it were reflective, and (3) that is not in general transitive.[5]

[5] It is important to make clear that the relation $We - Wp$ is very different from the relation S, between pairs of worlds w and points of view i, established by Hautamäki. Whereas the relation S is a set of pairs $<worlds, viewpoints>$, the relation $We - Wp$ is a set or pairs of worlds. As I said, according to Hautamäki, the weakest adequate way of defining S consists in saying that $<<w, i>, <w, i'>>$ belongs to S iff $V(p, w, i) = V(p, w, i')$ for some proposition p. With respect to some propositions p and some words w, a viewpoint i would be "alternative" to the viewpoint i'. Now, whereas the relation S, so defined, cannot fail to be reflexive and symmetrical, the relation $We - Wp$ may or may not be so. The worlds of emplacement do not have to be worlds in perspective, and the worlds in perspective do not have to be worlds of emplacement either.

3 Some Implications

I have introduced two general approaches to the structure of points of view. Now, I will examine the relationships between them discussing some relevant implications. There are many important ones with respect to the nature of points of view.

3.1 A Comparison Between the Two Approaches

There are many differences between the two approaches above presented. Whereas the first approach tries to analyse the internal structure of points of view, paying especial attention to their psychological ingredients in cases of points of view with a personal or psychological subject as bearer, the second approach focuses on the role of points of view as ways of having access to the world, and to ourselves, from some locations. According to the second approach, points of view have, in quite a direct way, a very strong relational character and a crucial modal dimension. Beyond other differences, this is so in all the authors analysed.

Paying attention to the internal structure of a point of view, the first approach is very sensitive to many distinctions that are far from the second approach. Some of the most important ones are the distinctions between personal/non-personal points of view, individual/collective points of view, conceptual/non-conceptual points of view, and possession conditions/attribution conditions.

However, in spite of these differences, the two approaches are compatible. There is no opposition between them. And here is no need to choose between them. Each approach places emphasis on different features of points of view. Furthermore, there are very important features that are present in both approaches. According to both of them, points of view are very peculiar entities with quite a strong relational character and with a not less strong modal dimension. As I have noted, this is directly so in the approach based on location and access. In the approach based on the model of propositional attitudes, things are not so direct. The relational nature and modal character of points of view are features derived from the relational and modal character of the conceptual and non-conceptual explicit contents of our mental states. But, in any case, there could not be points of view without those features.

3.2 The Relational Nature and Modal Dimension of Points of View

Points of view have an "internal structure", full of psychological ingredients in our particular personal case. And points of view also have a "role". That role consists in making accessible certain parts of reality from certain emplacements.

The role is defined by a complex relation between the subjects and reality. According to Moline, that relationship is essentially connected with a wide set of behavioural dispositions. According to Hautamäki, it has a structure that can be expressed in terms of modal logic. According to Moore, it entails a certain metaphysical conception about how our perspectives can be combined. In any case, we can ask: How can we understand the relationships between the two general aspects of a point of view, its internal structure and its role? More concretely, how can we understand the relationships between the psychological ingredients of a point of view, a part of its internal structure, and the role of that point of view?

The role of a point of view cannot be understood simply as a way of identifying and referring to the psychological ingredients included in its internal structure, neural states in our biological case. It cannot be understood in the same way in which Armstrong [1] and Lewis [9] understood the functional roles associated with mental concepts. According to these authors, mental concepts define some functional roles. And through those functional roles we were referring to the neural states that, in our particular biological case, are instantiating those very functional roles. However, the psychological ingredients of points of view cannot simply be understood as the physical realizers of the roles associated with those points of view, i.e., in our biological case as some neural states instantiating and making possible the roles of our points of view. The functional roles of points of view include many ingredients which have not a psychological character.

The role/occupant, or function/realizer, etc., distinctions have been very influential in philosophy of mind. For authors like Armstrong and Lewis, functional roles can only have a, let us say, heuristic or instrumental value. They allow us to identify and describe, in a second-order language, kinds of states that in themselves have an intrinsic physical nature. In the case of points of view, we cannot apply that strategy. If we were to apply to points of view the Armstrong-Lewis' strategy, then we would lose the peculiar relational nature and crucial modal dimension of points of view.

The role of points of view cannot have only a heuristic, or instrumental value concerning the identification and description of the subjective components of the points of view. Points of view involve relationships with the world. And not only actual relationships, but possible ones. And that relational nature and modal dimension is not determined by the subjective components of points of view.

The ontological status of points of view really is singular. There are important reasons to consider that points of view are not reducible either to subjectivity, or to information, or to physics.

Against first appearances, points of view are not something merely subjective. As I have argued, they cannot be reduced to our subjectivity. Points of view include many ingredients that are not merely subjective. This is directly so when points of view are understood according to the model of location and access. That approach emphasises the modal character of points of view. They are understood as possibilities of access. When points of view are taken in that way, they override any merely subjective approach. But the model based on propositional attitudes also suggests the non subjective nature of points of view. In some cases, points of view involve psychological, even personal, subjects and psychological attitudes. But, they always involve many ingredients which are not merely subjective. In particular, points of view always involve sets of non-conceptual contents or sets of conceptual contents, and it is not obvious that those things can have a merely subjective status.[6]

Points of view are not reducible to information either. And this is so in spite of any effort to shorten distances between the notion of information and the notion of point of view. Even if we adopt a very general and wide-ranging notion of information, according to which it is possible to speak of things as the non-conceptual information linked to analogical representations, and we assume that points of view have a certain informational structure, points of view cannot be reducible to information. They cannot be reducible to information because points of view include many components which are not reducible to information. The bearers of points of view and the relationships maintained between them and the explicit contents of the points of view are two ingredients that are clearly not reducible to information.

Finally, it is also difficult to see how points of view could be reducible to physics. The crucial problems have to do again with the irreducible relational character and modal dimension of points of view. Unless these features are shown to be completely reducible to physics, points of view would not be reducible to physics either.

Points of view are relational entities modally qualified. Both their relational nature and their modal qualification go beyond what can be found in the bearers of the point of view individualistically considered. This is especially important in the case of personal points of view. If the nature and existence of points of view is relational and modal in the above sense, then points of view cannot be reduced to the psychological ingredients present in subjects like us when we are adopting those points of view. More precisely, points of view would be so reducible only under the assumption that those psychological ingredients also have the relational and modal mode of existence of points of view. In any case, that relational and modal mode of existence would go far beyond what can be stated by any informational and physical description.

All of that suggests a crucial distinction between

(1) the subjects, and
(2) their points of view

and a correlative distinction between

[6] The discussions between externism and internism, both in philosophy of language and in philosophy of mind, are relevant here. According to externism, neural or psychological states, individualistically considered, do not determine the contents of our thoughts or languages. Internism rejects that thesis. I have tried to avoid those discussions. In any case, my claims about the relational nature and modal dimension of points of view entail a different argumentative line for externism and against internism.

(3) what can be “internal to the subjects”, and
(4) what can be “internal to their points of view”.

To have certain contents in perspective, to have certain non-conceptual or conceptual explicit contents, can be something “internal to the point of view of a subject” without being something “internal to the subject” individualistically considered. To take seriously the notion of point of view leads to a rejection of the claim that points of view are merely something internal to the subjects.[7] Points of view redefine the spaces of “objective possibilities” relevant for a subject. Points of view open spaces of objective possibilities that would not be available if the subject were not to take those points of view, or if the subject were to abandon them.

In sum. we can say that points of view are “modalized ways of being in relation to the world”. And that those “modalized ways of being in relation to the world” are not reducible to subjectivity, nor to information, nor to physics.

It is clear that I have been arguing for the non-fictional character of points of view. Now, if models entail the adoption of a point of view, or the integration of a number of points of view, my arguments will also apply to the non-fictional character of models. Models are not fictions. And a very important reason for claiming that models are not fictions is the following one: models involve points of view, and points of view are not fictions.

4 Conclusions

As I said at the beginning of the paper, the notion of points of view is important in many contexts. And it is particularly relevant in scientific and technical disciplines when our knowledge is guided by models. A model offers a certain point of view from which we get cognitive or practical advantages. Models offer peculiar points of view in order to see and conceptualise reality. And they offer peculiar points of view in order to act upon it.

I have argued that points of view have an irreducible relational nature. And that they also have a no less irreducible modal dimension. Insofar as points of view have such a non-reducible mode of existence, references to points of view in scientific and technical fields cannot be seen as having only a heuristic or

[7] My distinction between “internal/external to a point of view” and “internal/external to the subject that is taking that point of view” entails important consequences in the internism/externism debates about mental states. Farkas [4] makes a very good point rejecting the claim that the relevant internal/external frontier can be drawn in something like the body, or the brain. However, it is very different to draw that frontier regarding points of view than drawing it regarding the subjects that are the bearers of those points of view. If we were to distinguish between “internal/external to a point of view” and “internal/external to the subject that is taking that point of view”, then it could make sense to be internalist with respect to the point of view and externalist with respect to the subject taking that point of view.

pragmatical value, and models trying to incorporate those points of view cannot be seen either as constituting some kind of "second-class" discourse. If points of view are non-reducible parts of reality, then references to them will have to be simply unavoidable.

Does this entail a complete rejection of our scientific image of the world? There are two promising ways of elaborating a negative answer. The first one consists of emphasising the circumstantial character of points of view. Points of view usually include many indexical features that are very difficult to reduce to non-indexical ones. The ways of being in relation to the world, constitutive of points of view, cannot be reduced to relationships between non-relational ways of being if indexicality is excluded from those non-relational ways of being. From this perspective, the hard problem with points of view is not their supposed subjective character. As I have said, it is arguable that points of view are not subjective entities. The hard problem comes from their indexicality.

The second way of giving a negative answer to our question is even more radical than the first one. Again, it is linked to the modal dimension of points of view. The sort of efficacies achieved when a point of view is adopted, in particular through our personal points of view, suggest that both the bearers of the points of view and the things that are under the perspective of the points of view can change in ways that cannot be forecast in advance. The efficacies of points of view are essentially connected with "novelty". And this entails that we could not have any complete point of view about our possible points of view. From this perspective, if there is something radically irreducible about our points of view it is that sense of "openness". In particular, there would not be any point of view capable of accounting for all the possibilities of our epistemic, practical, normative and evaluative points of view.

Any of those lines of thought, solely or in combination, offers very strong reasons for claiming that the ways of being in relation to the world, that constitute our points of view, cannot be reduced to anything else. Points of view have to be aggregated to our scientific image of the world. Their mode of existence is not reducible to anything more basic or fundamental.[8]

References

1. Armstrong, D.: A Materialist Theory of the Mind. Routledge, London (1968)
2. Crane, T.: The nonconceptual content of experience. In Crane (ed.) The Contents of Experience. Cambridge University Press, Cambridge (1992)
3. Crane, T. (ed.): The Contents of Experience. Cambridge University Press, Cambridge (1992)
4. Farkas, K.: The Subject's Point of View. Oxford University Press, Oxford (2008)
5. Hales, S.: Review of Adrian Moore's book "Points of view". Mind **109**, 433 (2000)

[8] The non-fictional character of models has recently been emphasised by Magnani [10]. The arguments presented in my paper serve to complement some points of Magnani's approach.

6. Hautamäki, A.: The logic of viewpoints. Stud. Logica. **42**(2/3) (1983)
7. Hautamäki, A.: Points of view and their logical analysis. Acta Philosophica Fennica vol. **41** (1986)
8. Jackson, F.: What Mary didn't know. J. Philos. **83**, 291–295 (1986)
9. Lewis, D.: Mad pain and Martian pain. In: Block, N. (ed.) Readings in Philosophy of Psychology, vol. I. Harvard University Press, Cambridge (1980)
10. Magnani, L.: Scientific models are not fictions. Model-based science as epistemic warfare. In: Magnani, L., Ping, L. (eds.) Philosophy and Cognitive Science, Western and Eastern Studies, vol. 2, pp. 1–38. Springer, Heidelberg (2012)
11. Moline, J.: On points of view. Am. Philos. Q. **5**, 191–198 (1968)
12. Moore, A.: Points of view. Philos. Q. **5** (1987)
13. Moore, A.: Points of View. Oxford University Press, Oxford (1997)
14. Nagel, T.: The View from Nowhere. Oxford University Press, Oxford (1986)
15. Noë, A.: Is perspectival self-consciousness nonconceptual? Philos. Q. **52**(207), 185–194 (2002)
16. Noë, A.: Action in Perception. MIT Press, Cambridge (2004)
17. Peacocke, Ch.: Sense and Content. Oxford University Press, Oxford (1983)
18. Peacocke, C.: Analogue content. In: Proceedings of the Aristotelian Society, supplementary volume 15, pp. 1–17 (1986)
19. Peacocke, C.: Perceptual content. In: Almog, J., Perry, J., Wettstein, H. (eds.) Themes from Kaplan. Oxford University Press, Oxford (1989)
20. Peacocke, C.: A Study of Concepts. MIT Press, Cambridge (1992)
21. Peacocke, C.: Scenarios, Concepts, and Perception. In Crane, T. (ed.) The Contents of Experience. Cambridge University Press, Cambridge (1992)
22. Perry, J.: The Problem of the Essential Indexical. Noûs **13**, 3–21 (1979) [Also in Perr, J.: The Problem of the Essential Indexical and other Essays, pp. 33–50. Oxford University Press, Oxford (1993)]
23. Putnam, H.: Renewing Philosophy. Harvard University Press, Cambridge (1992)
24. Ryle, G.: The Concept of Mind. Chicago University Press, Chicago (1949)
25. Russell, B.: Problems of Philosophy. Oxford University press, London (1912)
26. Sellars, W.: Science, sense impressions, and sensa: a reply to Cornman. Review of Metaphysics **23**, 391–447 (1971)
27. Williams, B.: Descartes: The Project of Pure Inquiry, 320 pp. Harvester Press, Hassocks (1978)

The Arabic Script, from Visual Analogy to Model-Based Reasoning

Christian Tamas

Abstract For the Arabs in the beginning was the sound and the letter. Only afterward, the written word embodied and conditioned by the Qur'an. The phonetic and graphic shapes of the letters inspired a series of analogies which related to God as universal wholeness and script. Basing themselves on the close observation of the Qur'anic script as absolute matrix the Arab scholars began to model it constructing formal theories to describe and explain its meanings and their applicability in everyday life. Thus, the Qur'anic written text and the geometrical forms derived from it extended to all aspects of real life subliminally placing man into an iconic world of letters which intermediate between theories, applications and their absolute model.

1 Introduction

Analogy is a process of comparing similarities between two neither completely similar nor completely different concepts [1]. As such, analogy deals with two elements: the source and the target, where the source is represented by a known object, while the target is usually an unknown object. When the two objects are connected, the analogy between them results into a functional relation, where both the source and target behave or function in a similar way, a structural one, where the source and the target have similar appearances or structures, or into a combined functional and structural relation [2]. In other words, analogy is related to the alignment of the elements which represent the source and the target [3]. But this

C. Tamas (✉)
Alexandru Ioan Cuza University of Iasi (Romania), Iasi, Romania
e-mail: christiantamas@yahoo.co.uk

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_8, © Springer-Verlag Berlin Heidelberg 2014

kind of relation is present not only between objects, but also between the relations that take place between the objects and also between the relations between relations.

On the other hand an analogy is coherent if it has a coherent structure in terms of similarity, structure and purpose [4], becoming isomorphic when the analogy reaches its utmost structural consistency, that is when the source-target relations are identical and the connected elements possess a series of identical features. From a formal point of view, analogies can be verbal and visual. They are visual when the use of a concrete image analogous to an unknown object have a major communicative impact and as such, they become iconic when they stand alone or when accompanied by oral or written analogies.

When applied to the Islamic space traditionally accustomed to think in doublets according to the assertion made in the Qur'an stating that the material logos embodied in the form of the Qur'anic script is the perfect, physical copy of the metaphysical logos kept in heaven on the so-called "guarded Tablet",[1] the mechanisms of creating analogical processes become difficult to understand. And this difficulty resides mainly in the fact that the Islamic mind and way of thinking formed and developed in a space where the physical and metaphysical worlds remain tightly united even now in an almost symbiotic way. In other words, Islam is not a simple religion as it may seem to many of those narrowly anchored in the auto-sufficiency of other systems of thinking, but a way of living and thinking very different from our own, where nothing can be conceived outside the "roundness" or completeness of the Qur'an embodied God as absolute model.

Diagrammatic and pictorial representations are among mankind's oldest forms of communication. Peirce, for instance, affirmed that diagrammatic representations are "veridically iconic, naturally analogous to the thing represented" [5]. That's why, for him an algebraic equation was an icon, algebra was a kind of diagram, and language was a kind of algebra.

Starting from the *'alif* and the *sukūn,*[2] analogous through form and interpretation to the line and the circle in geometry, to one and zero in mathematics, to the beginning and the end as intrinsic characteristics of God, the Arabic script was conceived and seen as a complex diagrammatic construction characterized by the exacerbation of sign's iconicity that goes beyond its qualities of mere image, becoming a dynamic *representamen* as ultimate meaning which polarizes the efforts of human mind (Fig. 1).

Thus, the line, the point and the circle (seen as the contour of the point), as well as the ciphers "1" and "0" as representations of the logophanic Divinity shaped in the golden age of Islamic thinking a strange process by which the meaning and the shape of the Word were at least subliminally converted into a matrix of models applicable in every circumstance and field of real life.

[1] *The Qur'an* 85:22.

[2] In Arabic, graphic sign shaped as a small circle denoting the absence of a vowel.

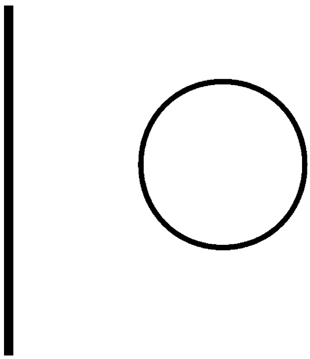


Fig. 1 The *''alif* and the *sukūn*

2 In the Beginning was the Letter

For the Arabs in the beginning was not the word, but the letter. This was the first physical representation of God who communicates with man.

In the Qur'an, the action of communication is an ambivalent one as God communicates, transmits a message but, at the same time, he communicates, transmits himself. Being the original Logos, he is at the same time the source and the sender of the message. The Qur'an is, therefore, a dual entity, being both the message and the author of the message.

As such, the Arabs considered the uttered Qur'an a verbal and acoustic analogy of God embodied in the Qur'anic message.

We have to do here with an analogy induced by the previous knowledge of God as creative Logos combined with the perception of sounds. In this respect, the Qur'an exercises a special attraction due to the fact that the correspondence between phonemes and the ideas expressed by words and sentences produces a profound impression. The musicality of the Qur'an as well as the alternation of slow and rapid rhythms generate powerful acoustic images able to deeply touch the receiver.

According to Bufano [6]

> Our physical world is full of elements, which have in themselves a significant value for the consciousness states of everyday life. The ticking of the clock is a sonorous stimulus that establishes a rhythm. In this respect, music can determine an amplification of the hypnotic answers whose mechanism is based on a Pavlovian-type conditioning. Thus the sonorous stimulus will gradually replace the verbal one, because music represents an ensemble of evocative stimuli, which intensely activate the emotive zones. Voices and sounds are also very used. The evident advantage is the freeing of the conscious mind from the task of giving instructions. In the first, inductive phase, one can use objects in order to attract the attention, such as lit or painted candles. Milton Erickson used in this respect a quartz crystal placed on a desk.

For the Arabs the lit candle becomes Book, the voice becomes recitation, and the sounds direct the attention towards the point that represents the *incipit* of the winding letter without beginning and without end.

Therefore, for the Muslim, the Qur’an is more than a religious message, it is a skill acquired through learning, an automatism permanently maintained by the use of the language and susceptible of being activated at the moment of listening or reading a Qur’anic passage.

3 The *’alif*, Matrix of the Arabic Script

Until the Qur’an, the Arabs did not have a well-defined writing system of their own. Many western scholars, such as Theodor Nöldeke in 1865 [7] and others who, years later, based their theories on the “al-Namarah” inscription [8], opined that the Arabs used an alphabet linked to or derived from the so-called Nabataean alphabet developed from Aramaic. Other scholars basing themselves on the classic Arab authors consider that the Arabic alphabet derives from the so-called *musnad* script [9] used in the ancient Southern Arabia. In spite of the divergent opinions both western and Muslim scholars agree that the Arabic script as we know it today developed out of the necessity of writing the Qur’anic words for safekeeping and learning, and above all for clearly deciphering and understanding their meanings. This need led not only to the construction of a precise writing system but also to a unitary language, the classic Arabic, whose purely consonantal structure determined the architecture of a logic almost mathematical system based on strict patterns derived from the verbal consonantal root [10].

Continuing the ancient tradition of pictorial analogies that can be traced back to the origins of the alphabet, the Muslim scholars developed a series of theories based on analogical-deductive processes centered on the point and the line as fundamental cosmological elements constantly related to the idea of divinity, which shaped and determined the sacred dimension of the Arabic script.

Thus, the *’alif*, the first letter of the alphabet, was considered an analogous representation of the Creator and as such the matrix of the entire alphabet.

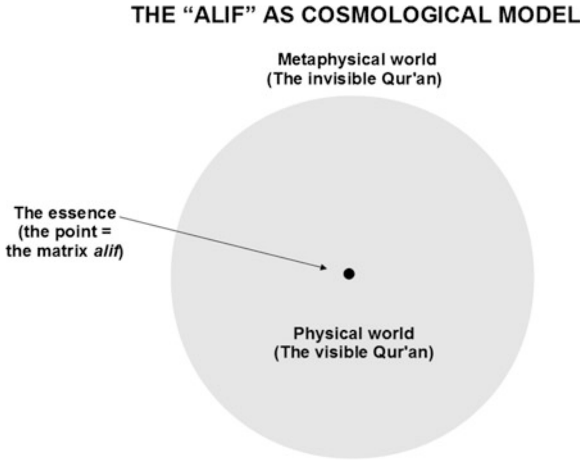
The *’alif* was seen as a point multiplied in order to form a line. Thus “the point creates the *‘alif* and and the *alif* all other letters, for while the point is the symbol of the Divine Ipseity, *alif* symbolizes the station of Oneness.” [11] As such it was seen as the creative principle of all letters contained within itself, the way God was the initial point of the universe contained within himself.

’Ibn ‘Aṭā’ (d. 1309) saw in the ’a*lif* the first man. Analogous, at the same time, to the metaphysical individuality and uniqueness of God “the *’alif* is, for the letters, like Adam, and *hamza*, the diacritical sign which came out of it, is like Eve. The twenty-eight letters are procreated by this *‘alif*.” [12].

And its characteristics are, in the opinion of the same author: (1) the rectitude *(qawām)*, (2) the axiality *(mihwar)*, (3) the verticality *(qā’im)*, (4) the balance *(mu‘tadil)*, and the elevation *(muntasib)*. For ’Al-Hallāj the*’alif* was a spiritual monad which contained all the others monads [13].

In the same line, Ibn ‘Arabī [14] used to opine that the*’alif*-ul, under its sonorous shape, represented a unique and distinct monad and not only because it

Fig. 2 The ’alif as cosmological model



was the first letter of the alphabet but because it was the nearest to the divine breath, to the creative spirit of God. According to the Muslim philosopher, the’*alif* did not originate from the other letters because while these ones could have been destroyed and then reconstructed out of the’*alif*, the’*alif* could not dissipate in their shapes (Fig. 2).

Therefore this consonant is present in every letter and word, the way the monad is present in everything belonging to this world taking into account that no matter what sound would be pronounced the’*alif* would necessarily open its way as beginning of the air flux through the larynx, constituting thus a revealing of existence.

Applying to the alphabet the theory according to which the number “1” contains within it all the other numbers, Ibn ‘Arabī considered that this letter had two forms or two dimensions: (1) a graphic shape made of a vertical line and (2) a sonorous form similar to the glottal stop *hamza*.

The’*alif* was the primordial letter, it was the first and all, the way God was the first and all. When pronounced, its sound is a pure and neat breath, as it originates from the divine Essence, which is its primordial matrix, and contains all the other articulations of Arabic, occlusives, gutturals, palatals and fricatives. Seen in the world-circle, the ‘*alif* is, from the arithmetic point of view the number “1”, from the geometrical point of view, a line, and from the calligraphic point of view, the circle’s diameter where the individualization of the other letters takes place.

The Muslim mystic Al-Jīlī (1365-ca.1424) related the’*alif* to the idea of harmony among human beings. According to him the first letter was named’*alif* because all the meanings derived from it express the approach between people the way its substance unites the letters between them [15].

This verifies also from a formal point of view, because, taking into account that they have geometric shapes, the letters can be reduced to the ’*alif*, a geometric shape itself. That’s why, for the cited author, when written or uttered, all the letters contain the ’*alif*.

In the same sense, Al-‘Alāwī explained the fact that the spatial forms of the letters represented only a transformation of the*’alif* which was their common element.

Moreover the *’alif* represented the first obvious manifestation of the point, as it appeared under a more incomparable than comparable form and maintained its existence in every letter, remaining distinct from them. That’s why, for Al-‘Alāwī [16] the*’alif* was a symbol of the unique God whose existence isn’t preceded by anything, it was a point, an absolute beginning. Like God, it was the first and the last, revealed and hidden at the same time. Before revealing itself the point was something impossible to anticipate, it was in a place where all the letters were hidden in its hidden essence.

In his primordial state, even if God has a conceptual presence, this presence is somewhat latent, somewhat not yet fully manifest. Precisely this primordial latent presence of the transcendental essence was considered in the 8th–10th centuries to be the essence of the point as primordial principle of all geometric shape.

The point was the signifier of unity, not the unity in itself. Therefore the point in itself but also as an a priori projection manifest under the form of the circle and line, the basic elements of the alphabet, was the ultimate determinant of the Arabic script.

In this sense, Eva Vitray-Meyerovitch [17] stated that “the divine Essence, in its absolute unity, is often symbolized, in the mystics’ language, by the letter *’alif*, a simple line lacking any diacritical sign […]. Precisely by virtue of the numeric symbolism, possible only in a language where the letters have arithmetical values, the *’alif* can be considered the archetype of the entire alphabet.”

As such, the *’alif*, matrix of all letters and, by extrapolation, of the entire Arabic script, was seen as the first form of the absolute Unity which emanates from the indefinable metaphysical point.

4 Iconic Functions of Calligraphy

In general, it is considered that language presents itself under the linear form of the enunciation which is produced by the sequence of words in the sentence, phrase and discourse. In the case of the visual communication of pictorial type linearity is no longer possible, because of the two dimensions involved and reinforced by the fact that the pictorial image is prospective. We deal therefore with an analogical, non-linguistic language.

In the Qur’an, the linearity of the sacred enunciation is transformed into pictorial image, the analogical language being determined by the discursive nature of the enunciation that constitutes the nucleus and the reason to be of the image.

The sacredness of the Arabic language as vehicle/channel of the divine Word, had a role in the spatial fixing of the message, the way the oral language allowed its fixing and spreading over time. The Qur’anic message confers to the calligraphic script (designated, in Arabic, by the word *khatt*) their maximum hieratic dimension.

Calligraphy is a visual art, but in Islam the visual nature of the letter became consubstantial to its metaphysical model. Thus, calligraphy was seen as a spiritual geometry built with material instruments [18].

The sacred image of God is contained in the iconography of the letter, and the formal aspect, as a key element of the script, transfers to the calligram an important role in preserving and transmitting the divine Word. Being about God, the word should be spread in a way able to send to the Revelation miracle. The transcendental basis of the script contained in its own shape confers to the ornamental writing a full calligramatic character, since in it we can find the Essence of the word itself that is not a mere vehicle of thought [19].

Taking into account its plastic, almost sensual meaning, some Muslim scholars saw a connection between calligraphy and erotic language, the fusion of the two conferring to the calligram the idea of totality and complexity required to represent the mystery.

Given its theophanic function, of transfiguration of the Logos, calligraphy has the mission to perfectly reflect the Word-God, the letter thus becoming a figurative representation of the Deity and the first fundamental key of interpretation offered to the uninitiated.

Thus, in the Islamic world the pure aesthetic value of calligraphy cannot be separated from the meaning of the word; this represented the reference axis of the calligraphers' work, who by the late Middle Ages, had already exhausted the last possibilities offered by the letter.

Calligraphy in Islam has the function of transmitting knowledge, it has the power to separate ideas, to recompose and preserve them. Ibn Khaldūn [20], supporting the communication function of calligraphy, stated that this one represented the word's first level of signification, being the first clear expression of the ideas contained in the soul and mind.

Another function attributed to calligraphy in Islam is that of being the translator of thoughts and feelings; as such it is often compared with the work of goldsmiths and weavers, because letters and lines interlink in a cosmic hermeneutics.

5 Model-Based Reasoning in Islam

When dealing with artificial intelligence, model-based reasoning takes into account inferential methods used in systems based on a model of the physical world. Thus, the main focus of application development is directed toward developing a model whose knowledge combined with observed data makes the artificial system derive conclusions.

The same is verifiable in the case of Islam, where the Qur'an acts as ultimate model of both physical and metaphysical worlds taking into account that the

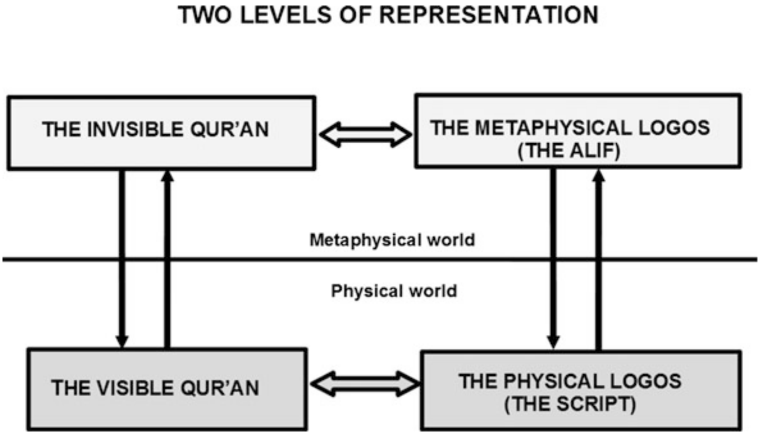

TWO LEVELS OF REPRESENTATION
THE INVISIBLE QUR'AN
THE METAPHYSICAL LOGOS (THE ALIF)
Metaphysical world
Physical world
THE VISIBLE QUR'AN
THE PHYSICAL LOGOS (THE SCRIPT)

Fig. 3 The Qur'an. Two levels of representation

physical Qur'an is considered a perfect copy or transcription of the metaphysical one.[3]

According to the Islamic conception, one of the forms of guidance provided by God to man is giving him the capacity to discern between good and evil. In the Qur'anic message, this guidance is represented by the signs, understood as natural realities, which become objects of the verbal signs communicated as revelations (Fig. 3).

Therefore, the revelations are seen as an authentic source of information, fact that avoids putting them into question. Thus, the revelation is for the Muslims an ultimate truth in the sense that any conclusion based on other sources must be filtered through the divine revelation, before being expressed as truth.[4]

In terms of practice, basing themselves on the close observation of the Qur'an as absolute matrix, the Arab scholars began to construct formal theories in order to define as accurate as possible its meanings and their applicability in everyday life by developing a grammar system and a specific science named *tafsīr*[5] which tried to establish patterns in all the fields related to man's life and activity mainly on the basis of deduction and logical analogy.

Usually defined by Abu Hayyan Andulasi and other Muslim scholars as "the science that discusses about the utterance of the Qur'anic words, about their meanings and connotations" [21], this method, calling for thorough knowledge of linguistics, philosophy, theology and jurisprudence, will be addressed, according to the Muslim terminology, from the following perspectives: (1) traditional, (2) rational, (3) scientific and (4) mystical, the first three using semantic-pragmatic

[3] Al-Qur'an, 43:4, 85:21–22.

[4] That's why the Qur'an is known as well under the name of *al-furqān* ("what distinguishes truth from untruth").

[5] Term apparently derived from the 2nd form root *fassara*, meaning "to open" or "to expose".

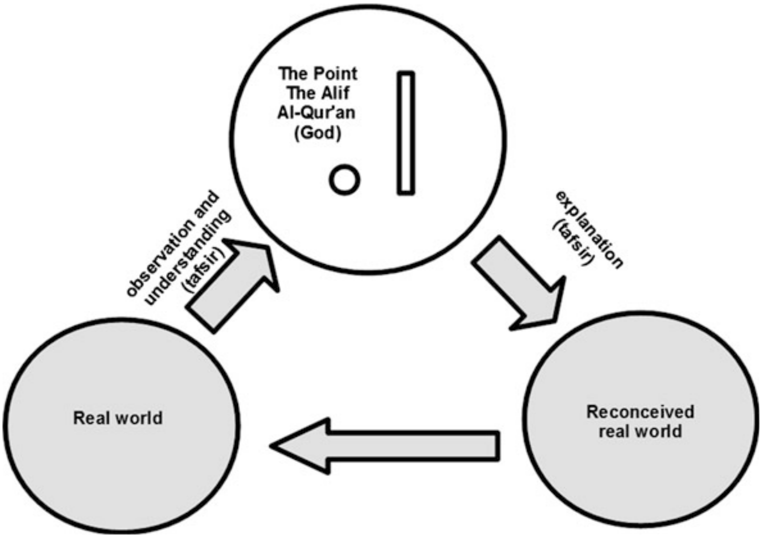


Fig. 4 Recreating real world through the Qur'an

methods of analysis, while the last one used the hermeneutic-semiotic method.[6] The *tafsīr* was used for constructing the Qur'an as physical model through specific scientific methods which allowed the building of meanings through the vocalization of the consonantal skeleton of the linguistic units and their insertion in an adequate context. After the completion of this task the *tafsīr* was used to re-conceive by interpretation, through the intermediary of the Qur'an as ultimate model, real-world situations which influence the perception on real world situations. In other words, any real situation must be confronted with the patterns established in accordance with the Qur'an, before being accepted or rejected as true or false (Fig. 4).

The Islamic world is generally a world governed by the physical Qur'an subliminally present in the Arabic script and circumscribed to the metaphysical one. Thus, although in Islam too both inquiry and application as basic practices of model-based reasoning cross the border between verified observations and the models used in order to explain facts, they are unthinkable outside the roundness of the cosmic Qur'an, their ultimate point of reference (Fig. 5).

Thus, the Qur'an and the geometrical forms derived from it and extended to all aspects of real life place man into an iconic world of letters which intermediate between theories, applications and their absolute model. The Arabic script seems thus very similar to the pictures of structures used by many model theorists to think about structures. If, when reasoning with diagrams, pictures or diagrams are

[6] This method is also known as *ta'wīl*, literally meaning the "return of the meaning to its original matrix".

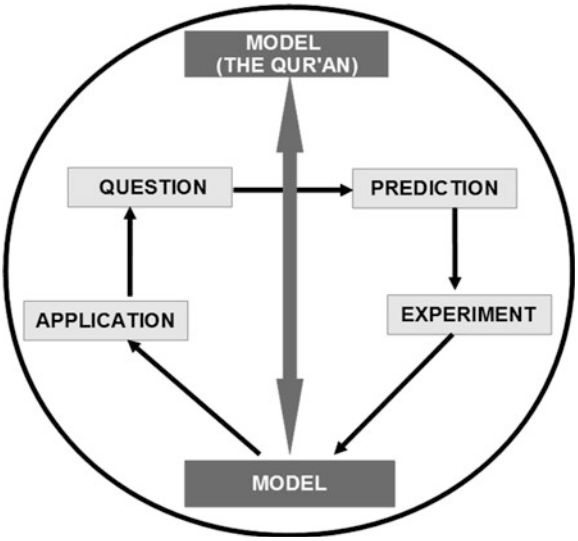


Fig. 5 The inquiry circle in Islam (Graph adapted from White, Shimoda and Frederiksen [22])

seen more as a form of language rather than as a form of structure, in Islam they are both, as language is besides its intrinsic quality a form of structure.

References

1. Orgill, M.K., Bodner, G.M.: An analysis of the effectiveness of analogy use in college-level biochemistry textbooks. J. Res.Sci. Teach. **43**(10), 1040–1060 (2006)
2. Curtis, R.V., Reigeluth, C.M.: The use of analogies in written text. Instr. Sci. **13**, 99–117 (1984)
3. Gentner, D.: Structure mapping. A theoretical framework for analogy. Cogn. Sci. **7**(2), 155–170 (1983)
4. Holyoak, K.J., Thagard, P.: The analogical mind. Am. Psychol. **52**(1), 35–44 (1997)
5. Pomorska, K., Rudy, S., Jakobson, R.: Language in Literature, p. 419. Harvard University Press, Cambridge (1987)
6. Bufano, A.: Autoipnosi e terapia, Vertici Network di Psicologia e Scienze Affini, http://www.vertici.com
7. Grohman, A.: Arabische Paläographie. Wien, Böhlau in Commission, p. 11 (1971)
8. Healey, J.F.: The Early Alphabet, p. 54. University of California Press, Berkeley (1990)
9. Abulhab, S.D.: DeArabizing Arabia. Tracing Western Scholarship on the History of the Arabs and Arabic Language and Script, pp. 233–235. Blautopf Publishing, NY-Ulm (2011)
10. Holes, C.: Modern Arabic. Structures, Functions and Varieties. pp. 99–100. Georgetown University Press, Georgetown (2004)
11. Nasr, S.H.: Islamic Art & Spirituality. p. 32. Golgonooza Press, Ipswich (1990)
12. 'Ibn ʻAṭā' ' Allah: Traité sur le nom ' Allah, (introduction, traduction et notes par Maurice Gloton), pp. 137–138. Les Deux Océans, Paris (2001)
13. Massignon, L.: Essai sur les origines du lexique technique de la mystique musulmane.In: Vrin, J. (ed.), p. 38. Paris (1954)
14. Yousef, M.H.: Ibn Arabi—Time and Cosmology, p. 155. Routledge, New York (2008)

15. Al-Jīlī, A.K.: Al-kahf wa ar-raqīm fī sharh bi'smi-llah-i-rrahman ar-rahīm. p. 20. Dar al-Kutub al-'Ilmiyah, Bayrut (2004)
16. Lings, M.: A Sufi saint of the twentieth century: Shaikh Ahmad Al-'Alāwī, his spiritual heritage and legacy. p. 153. University of California Press, Berkeley (1971)
17. Vitray-Meyerovitch, E. de: Mystique et Poésie en Islam, p. 171. Desclée de Brouwer, Paris (1982)
18. Derman, M.U.: The art of calligraphy in Islam. In: Ihsanoglu, E. (ed.) The Different Aspects of Islamic Culture, Culture and Learning in Islam, vol. 5, p. 570. UNESCO Publishing, Paris (2003)
19. Khatibi, A.: La blessure du nom propre. Denoel, Paris (1974)
20. Ibn Khaldūn: The Maqaddimah: An Introduction to History. (translation by Franz Rosenthal). vol. 1, Pantheon Books, New York (1958)
21. Ahmed, H.: Introducing the Quran: How to Study and Understand the Quran, p. 170. Goodword Books, New Delhi (2004)
22. White, B.Y., Shimoda, T.A., Frederiksen, J.R.: Enabling students to construct theories of collaborative inquiry and reflective learning: computer support for metacognitive development. Int. J. Artif. Intell. Educ. **10**, 151–182 (1999)

Mechanism and Phenomenon of Consciousness

On Models and Ontology in Dennett and Edelman

Paolo Pecere

Abstract The neurological explanation of consciousness has become in the last decades a widespread field of research among neurobiologists and philosophers of mind. The development of experimental models of consciousness involves a parallel search for a suitable ontological background. Although most researchers share anti-dualistic and naturalistic ideas, there are controversial claims about the ontological interpretation of phenomenological data. After sketching some historical premises of this issue, the paper focuses on two case studies: Dennett's "multi-draft" model of consciousness, and Edelman's theory of consciousness, included in his "theory of the selection of neuronal groups". Edelman's theory turns out to provide a better solution to the open issues of contemporary research, since it avoids speculative hypotheses and dismissive attitudes, while leaving room for experimental and conceptual developments in a classical, "Newtonian" methodological style.

I will present some remarks about the use of mechanistic models in contemporary neurosciences and its ontological implications, focusing on two case studies: Dennett's radical program of a materialistic "explanation" of consciousness and Edelman's interpretation of his own neuroscientific model of consciousness. The discussion is best introduced by means of some introductory remarks about the Cartesian legacy in neuroscience, since contemporary issues about the neuroscientific explanation of consciousness—has it has been often recognized—still owe much to a Cartesian philosophical background.

P. Pecere (✉)
Università di Cassino e del Lazio Meridionale, Cassino, Italy
e-mail: paolopatch@yahoo.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_9, © Springer-Verlag Berlin Heidelberg 2014

1 Mechanistic Models and Dualistic Metaphysics: A Cartesian Controversy and its Legacy

« I suppose the body to be nothing other than a statue or machine made of earth » ([1], XI, 120): this famous statement made by Descartes in his treatise *L'homme* (first published in 1662) largely influenced the study of the brain by connecting mechanistic physics with anatomical analysis. After seeing Descartes' book fresh off the press, Steno wrote in his *Discours sur l'anatomie du cerveau* (1665): « since the brain is a machine, we have no reason to hope to discover its design through means any different from those used for discovering the design of other machines. The only thing to do is what we would do with other machines, taking apart its components piece by piece and considering what they can do, separately and together » ([22], pp. 32–33). The use of a mechanistic model of the brain, of course, was limited in Cartesian philosophy by the metaphysical distinction between the essence of soul and the essence of body, which excluded the very possibility of explaining the higher mental faculties, and consciousness itself, by means of mechanistic physics. This claim was highly appreciated by thinkers such as Malebranche and Leibniz: the latter wrote, in a famous page of the *Monadology*, that « perception, and anything that depends on it, cannot be explained in terms of mechanistic causation » [est inexplicable par des raisons mecaniques], arguing that visiting the interior of a machine « would show you the working parts pushing each other, but never anything which would explain a perception » ([19], § 17, p. 609). Since the XVIIth century this claim has been contested by many thinkers—such as Spinoza and La Mettrie—who underscored the heuristic power of the mechanistic models for the understanding of the mind and presented dualism as a metaphysical prejudice and an impediment to scientific inquiry.

An anti-dualistic—and therefore anti-Cartesian—perspective has gained renewed attention in neurosciences of the second half of the twentieth century, as the developments in biology and the new techniques of brain-imaging have led to different attempts to explain the "mechanism" of consciousness, without resorting to dualistic ontological hypotheses on the mind. Anti-Cartesian chapters, in particular, are one of the common features of the main books on the theory of consciousness in the last 20 years, authored by both philosophers and neuroscientists who developed mathematical and/or mechanistic models of consciousness and its different properties: think of Gerald Edelman and Antonio Damasio, Patricia Churchland and Daniel Dennett. A different, sympathetic judgment about Descartes's legacy has been formulated by French leading neuroscientist Jean-Pierre Changeux, who considers Descartes as a major forerunner of any successive physical explanation of the brain functions. Changeux adheres to the old fashioned—yet quite questionable—historiographical idea that Spinoza and La Mettrie represent the straightforward development of Descartes' mechanistic program, whose implications were materialistic from the outset ([6], pp. 47–54). Changeux himself claimed that the hypothesis of a physical explanation of the mind (including consciousness) is the only heuristically positive option for neuroscientific research. As he put it in his programmatic book *L'homme neuronale*, there is no way for neuroscience but to assault the « Bastille of mind » ([5], p. 210).

In spite of any polemical accent, it should be recognized that both (a) the idea of mind and (b) the idea of mechanical explanation, that form the background for contemporary anti-Cartesian programs, owe much to the metaphysical foundation of modern science of nature provided by Cartesian philosophy.[1]

(a) The contemporary « problem of consciousness », as the quest for the neurobiological explanation of the most general qualitative feature of experience, usually presupposes Descartes' identification of the mind with « thought » , considered, in turn, as « everything which takes place in us so that we are conscious of it » ([1], VIII, p. 7). This problem, to be sure, would make no sense on the background of—say—Aristotelian hylemorphism, with its threefold soul as the "substantial form" of life and thinking, for here there is no gap to be further explained between matter and mind. Indeed, Descartes' list of biological phenomena than can be explained by means of a purely mechanistic account (see [1], XI, p. 202; [1], VII, pp. 229–230), includes all the functions of peripatetic vegetative and sensitive soul, which, in turn, correspond to the contemporary « neurological unconscious » , as the set of genetically or empirically stored abilities. Descartes' « mind » corresponds on the other hand to the conscious sensitive, imaginative and intellectual perceptions, which are precisely the phenomena investigated by contemporary theories of consciousness.

(b) The explanatory models developed by contemporary neuroscience are grounded on neurons and their physico-chemical activities, and as such they reflect the metaphysical distinction of matter (as lifeless extension) from mind operated by Descartes, while rejecting at the same time the very existence of a separate immaterial soul. This produces the need for an alternative explanation of consciousness and voluntary activity.[2]

On the whole, one can say that Cartesian ideas of matter and mind are essential for the very formulation of research programs in contemporary brain and cognitive sciences. This is still true today, as this conceptual heritage is more and more acknowledged and considered controversial by several leading researchers in both philosophy of mind and neurosciences: to dispose of mind-matter dualism is

[1] The presence of Cartesian ideas in philosophy of mind and neuroscience has been noticed several times in the twentieth century. This has been often considered as a starting point for philosophical criticism, which has been advanced by quite different perspectives (just think of Ryle and Heidegger). Most recently the dependence of twentieth century neuroscience on Cartesian dualism has been investigated by Maxwell Bennett and Peter Hacker. These authors consider the Cartesian attribution of mental properties to the soul as the exemplar model of the "mereological fallacy" of attributing mental faculties to the brain, which would be widely present in neuroscience ([2], pp. 43–44, 68, 160–161). On this part-whole problem see § 5 below.

[2] According to Descartes, we have « clear and distinct » ideas of both the separate existence of the immaterial soul and the action of the soul on the body. The latter's evidence therefore is not to be disputed or further analyzed. For a penetrating account see Garber [18], pp. 168–188.

indeed an ontological ideal that plays a crucial role in recent research on the neural correlates of consciousness.

In front of this complex Cartesian legacy, which here I cannot examine more in detail,[3] a series of questions arises: can the models of consciousness developed in contemporary neurosciences be considered as steps towards a reduction of consciousness to a mechanical process (that is, an ontological reduction of consciousness to matter)? Or do they play a heuristic role in the search for a ontologically different, non-materialistic theory? Which is, on the other hand, the methodological role of the very phenomenological evidence about conscious thinking that led Descartes to postulate metaphysical dualism? In contemporary philosophy and neurosciences these issues turn out to be controversial, as I will try to show by analyzing two different and contrasting cases.

2 Dennett: Mechanical Hypothesis and Explanation of Consciousness

Daniel Dennett, in his book *Consciousness explained* [9], sets out an explanatory hypothesis about consciousness which is connected to a « naturalistic-mechanical » ontology. He presents his view as opposed to the « reactionary » claim of those (such as Noam Chomsky, Thomas Nagel, Colin McGinn) who deny the possibility of a naturalistic-mechanical explanation of consciousness, and traces this view back to the Cartesian Age and to Leibniz's mistaken conflation of an epistemic problem with an ontological judgment.[4] Dennett's commitment to the naturalistic program is presented as a heuristic consequence of the fact that mind–body dualism is an impediment to scientific research and encourages the anti-scientific claim that consciousness is a « mystery ». According to Dennett, indeed, the introduction of mental properties in the description of conscious processes provides no theory at all, or, as he pungently puts it: « accepting dualism is giving up ».[5]

[3] It must be observed that Descartes himself devoted a substantial part of his work to discussing the unity of body and soul in its metaphysical, medical and ethical aspects. This fact is usually not recognized in contemporary criticism against Cartesian dualism. For an overview of this aspect of Descartes' philosophy, including an appraisal of its seminal role in grounding psycho-physical explanations in medicine, see Voss [29], pp. 186–196.

[4] For the latter argument see Dennett [11], pp. 1–10. I will elaborate on this problem in § 5 below.

[5] Dennett [9], p. 37: « There is a lurking suspicion that the most attractive feature of mind stuff is its promise of being so mysterious that it keeps science at bay forever. This fundamentally antiscientific stance of dualism is, to my mind, its most disqualifying feature, and it is the reason why in this book I adopt the apparently dogmatic rule that dualism is to be avoided at all costs. It is not that I think I can give a knock-down proof that dualism, in all its forms, if false or incoherent, but that, given the way dualism wallows in mystery, accepting dualism is giving up ».

Dennett therefore supports an explanatory theory of consciousness « within the framework of contemporary physical science » ([9], p. 40). By contemporary physical science Dennett considers standard physical science, and does not consider speculative conjectures such as Penrose's and Chalmers' about the possibility of new physical theories. Therefore, Dennett's program is to eliminate « mind-stuff » and to explain consciousness as a product of normally considered biophysical processes. Such a program, of course, is not altogether new: Dennett recasts reductionist and physicalistic ideas of twentieth century philosophy of mind, drawing on the most recent tools of cognitive science and connectionism.[6]

According to Dennett « human consciousness is itself a huge complex of memes (or more exactly, meme-effects in brains) that can best be understood as the operation of a "von Neumannesque" virtual machine implemented in the parallel architecture of the brain that was not designed for any such activities » ([9], p. 210). Dennett's hypothesis is constructed by drawing on contemporary developments in computer science (Von Neumann), linguistics (Levelt) and evolutionary biology (Dawkins). It considers consciousness as a « virtual machine » implemented in the brain; its ability to represent and express meanings is not subject to a central control (a central « Meaner »), but rather depends on a subconscious competition of « multiple drafts » , produced by parallel processes in different regions of the brain, whose resolution, that eventually leads to speech acts, depends on pragmatic criteria. The communication of meanings, in turn, corresponds to the ability to share « memes » in cultural networks, one of them being the very idea of the Self.

One of the most striking aspects of Dennett's hypothesis is the ontological denial of qualia, which depends on the philosophical criticism of the illusory contents of phenomenology. Contrasting the very idea of a quale, as a supposedly irreducible element of experience, Dennett argues that this can indeed be analyzed and "explained away" in terms of information and belief, and therefore it only exists in a fictional sense, rather than in a natural sense:

> « Heterophenomenological objects—i.e. qualia—are, like centers of gravity or the Equator, abstracta, not concreta. They are not idle fantasies, but hardworking theorist's fictions » ([9], pp. 95–96)
>
> «The heterophenomenology exists—just as uncontroversially as novels and other fictions exist. People undoubtedly do believe they have mental images, pains, perceptual experiences, and all the rest, and these facts—the facts about what people believe, and report when they express their beliefs—are phenomena any scientific theory of the mind must account for » ([9], p. 98).

[6] There are already several introductions to Dennett's theory of consciousness. For an overview see Schneider [25] and the brief critical assessments by Andrew Brook and Paul Churchland in Brook/Ross [4], pp. 41–63, 64–80. For a useful historical survey of physicalistic and anti-physicalistic trends in twentieth century philosophy of mind see Moravia [21], which does not cover contemporary naturalism such as Dennett's. For a more up to date account on contemporary issues see Velmans/Schneider [28] and McLaughlin [20].

Dennett rejects the label of “eliminativism”, insisting that his point here is to reconsider mind without mystery.[7] Given this important clarification, one still has to recognize that Dennett eliminates mind as a separate property or substance, reducing it to a fiction and an object of belief. But this does not mean that Dennett only wants to construe consciousness as being a matter of language: belief itself is a material brain process and therefore consciousness is indirectly inserted in a physical background. This crucial point can be highlighted by considering how Dennett’s program is rooted in Ryle’s and Wittgenstein’s philosophy of mind, and at the same time connects the latter’s criticism of mind to a brand new materialistic *pars construens*: Dennett’s ontological commitment with the machine model strictly depends on the philosophical project to connect a Wittgensteinian anti-metaphysical criticism of language, and in particular of private feelings, with an evolutionary and materialistic theory of mind framed within the tradition of computational cognitive science.[8]

In front of this bold theoretical claim, it is interesting to observe that what Dennett presents is « just the beginning of an explanation » ([9], p. 455). To be sure, according to Dennett the task of philosophy is to show whether such a theory is possible or impossible ([9], p. 41). Therefore he defends the possibility of a non-dualistic hypothesis, without entering the details of its realization. « All I have done, really, is to replace one family of metaphors and images with another […] It’s just a war of metaphors, you say—but metaphors are not “just” metaphors; metaphors are the tools of thought » ([9], p. 455). One may wonder, then, whether the materialistic reduction of consciousness in terms of material processes can be construed as a heuristic maxim, rather than as a fully grounded ontological commitment.

3 Edelman’s TSNG and the Role of Phenomenal Consciousness

It is very instructive to compare Dennett’s variously and strongly philosophically oriented conjecture with Edelman’s theory of the selection of neuronal groups (TSNG), which presents a quite different interpretation of the role of phenomenology in the scientific description of consciousness. Edelman’s theory—first fully articulated in *Remembered Present* [12] and later in a number of books such as *A*

[7] « Am I an eliminativist? I am a deflationist. The idea is to chip the phenomena of the mind down to size, undoing the work of inflationists who actively desider to impress upon themselves and everybody else just how supercalifragilisticexpialidocious consciousness is, so that they can maintain, with a straight face, their favourite doctrine: The Mind is a Mystery Beyond All Understanding » ([10], pp. 369–370).

[8] On this point it is very instructive to consider Dennett’s critical exchange with Maxwell Bennett and Peter Hacker, who defend a different development of Wittgenstein’s ideas ([3], for Dennett’s view see in part. pp. 77–89).

Universe of Consciousness [16]—includes probably the most elaborated evolutionary and anti-dualistic theory of consciousness in contemporary neuroscience. It is grounded on three empirical principles[9]:

(1) **Developmental selection**, as the formation of the gross anatomy of the brain, which is partly controlled by genetic factors, but involves a high degree of individual variation in the neural connectivity;
(2) **Experiential selection**, a continuous process of synaptic selection, occuring within the diverse repertoires of neuronal groups. This process may strengthen or weaken the connections among groups of neurons and it is constrained by value signals that arise from the activity of the ascending systems of the brain, which are continually modified by successful output;
(3) **Reentry**. The ongoing recursive dynamic interchange of signals that occurs in parallel among connected brain areas, and which continuously coordinates in time and space the activity of their maps. Edelman considers a massive presence of reentry as a distinctive feature of human brain.

On this background Edelman develops his hypothesis about the neural correlates of consciousness. In Edelman's model, consciousness depends at any given moment on the activity of different and distributed groups of neurons, which form the so-called « dynamical nucleus ». The dynamical nucleus is defined as a « functional cluster » of neurons, connected by reentrant interactions. Consciousness, on the other hand, is defined as the « ability to construct a scene » and operate multidimensional « discriminations » inside this scene.

The task of the theory, now, is to explain the emergence of consciousness, as it is phenomenally given: this includes properties such as unity, qualitativity, temporal ordering, intentionality ([14], pp. 119–120) Edelman argues that the phenomenology of consciousness can be connected with the underlying neural processes by means of different features of the latter's integration and differentiation of information. The constantly changing and integrating components of the dynamical nucleus, for instance, « correspond » to the changing contents of conscious experience and their temporal ordering: the dynamical connection of « value-category memory » and « perceptual categorization » first produced consciousness as a « remembered present » ([14], p. 55). The selective integration of different cortical maps accounts for the constructive aspect of consciousness (closure, filling of gaps, Gestalt effects). A quantitative measure of functional integration and differentiation helps to connect this hypothesis with mathematical models and to design experimental tests. The multifarious afference of sensory information and its mnemonic modulation account for the rich qualitative contents of experience.

[9] See Edelman [14], pp. 39–41. I will consider here the most recent expositions of the theory (starting from Edelman/Tononi 2001), which probably take into account some philosophical criticism of previous expositions. See below note 10.

In Edelman's model, on the whole, properties of the neural network correspond to phenomenal properties: here lies their "explicative" value. Edelman asserts very clearly that his theory (and any explanatory theory in general) cannot ever reproduce qualitative experience and that postulating, in this sense, the reduction of conscious experience to neural activity is a « category mistake ».[10] This is indeed a crucial point of his theory, which must be carefully analyzed. One has to separate the impossibility to reduce and therefore « eliminate » conscious experiences by means of neurological description from the « methodological inability » of present neuroscience to provide such a description, which is theoretically possible.[11] Even if we had a perfect neurological description of the immensely complex neural interactions (that is, to put it in Searle's terms, even if we knew « exactly how » reentrant mechanisms cause conscious states: see footnote n. 10), then we would not have « eliminated » or « reduced » (or, to put it in Dennett's terms, « explained away ») the conscious experience, as the natural way that enables us to be informed about our interaction with the world.

4 Phenomenon and Mechanism of Consciousness: A Comparison Between Dennett and Edelman

The exposition of Edelman's theory has already introduced the strong difference between his own and Dennett's program. First, Edelman's elaboration on selection theory is considerably different from Dennett's. Whereas in Dennett selection occurs among possible speech acts, on the "software" level, in Edelman the selection of the fittest populations of neurons is a fundamental feature of the brain connectivity and results in the plasticity of the neural architecture itself: in human

[10] Edelman [14], p. 125. This is possibly a reply to critical remarks advanced by John Searle with regference to Edelman's previous books *Remembered Present* and *Bright Air, Brilliant Fire* [12, 13]. Searle considers Edelman's theory as « the most thorough and profound attempt that I have seen in the neurobiological literature to deal with the problem of consciousness ». Nonetheless, he considers Edelman's theory unsatisfactory, because it does not explain how qualia are produced by the neural activity: « Assuming that we understand how the reentrant mechanisms cause the brain to develop unconscious categories corresponding to its stimulus inputs, how exactly do the reentrant mechanisms also cause states of awareness? One might argue that any brain sufficiently rich to have all this apparatus in operation would necessarily have to be conscious. But for such a causal hypothesis the same question remains—how does it cause consciousness? And is it really the case that brains that have these mechanisms are conscious and those that do not are not? So the mystery remains » ([26], pp. 48, 50). Searle's essay is a revised version of a review in « The New York Book Review », November 16, 1995.

[11] See e.g. Edelman [15], p. 145: « Indeed, at present, because we lack the means of fully detailing the hyperastronomical interactions of core neurons, C [the conscious system] provides the only indicator we have of any overall core state, C' [the neural system]. Indeed, our methodological inability to reduce to cellular or molecular terms the mental or conscious events accompanying fields such as ethics and aesthetics that emerge when we speak "C language" to each other should not be construed as arising from the existence of some radically inaccessible domain ».

brains, the “hardware” level cannot be separated from the “software” level. Second, Edelman is more cautious about the possibility of neurological explanations of single phenomenological data. The attempt to develop a physico-mathematical model of the brain processes results in the admission that the only evidence that can be mastered of such a complex physical system is—at present—statistical: the dynamical nucleus is defined by means of a measure of « neural complexity », grounded on the statistical theory of information.[12] But there is no evidence, in Edelman’s works, that a more advanced theory will be able to provide a more finely grained mechanical description of single qualia, for the latter are inserted in the unitary and multidimensional conscious scene corresponding to the dynamical nucleus.

On the whole, though Edelman intends to « complete Darwin’s program » ([14], pp. 1–3) by naturalizing mind in terms of biological evolution, his global biological approach provides a quite different way to naturalism than Dennett’s. Two global features of the nervous system—plasticity of the brain and complexity of neural interactions—support Edelman’s conclusion that phenomenal consciousness is a unique means of understanding the human mind, that was developed in the evolution of the human organism in order to represent the individually different and highly complex brain processes. Edelman considers this conclusion to be in direct opposition to the computer science model of the brain, and this rebuttal implies the joint rejection of any eliminationist program in neuroscience, including those that are built on computational metaphors. Contrary to the computational model, the phenomenal content of mind is not the single-channel output reduction of a parallel process of elaboration of data (or even, as Dennett puts it, a misleading construct of folk beliefs); phenomenal consciousness peculiarly expresses « complex discriminations » produced by the parallel activity of the brain, in order to put them at work by interacting with the ambient: « qualia » —as Edelman repeatedly underscores— « are these discriminations » , and therefore they exist ([14], p. 70). Indeed, the very « logical » model of mind, as a set of rules designed to perfectly decipher codified sense-data, is opposed by Edelman to his own interpretation of the fundamental process of mind, the « recognition of configurations ».

To be sure, Dennett and Edelman share some crucial views about neuroscience. They both consider a « mechanistic » model as heuristically fundamental for the sake of a scientific theory of consciousness, and support an ultimately naturalistic ontology, at least insofar as consciousness is identified with a neural « process » and does not require the position of any immaterial being. Both Dennett and Edelman, moreover, are active in the field of A.I. and consider the development of thinking artifacts as a crucial enterprise in order best to understand and possibly to

[12] This theory is presented in Edelman/Tononi [16], pp. 125–138. Since this section of the book contains mainly Edelman’s and Tononi’s technical work on the measurement problem I do not analyze it in details. See Tononi [23].

reproduce human thinking.[13] Nonetheless there are philosophically crucial differences in the way mechanism and phenomenon are related in their different theories, which we can summarize by distinguishing the (a) ontological from the (b) methodological point of view:

(a) According to Dennett, the phenomenon of consciousness (as the representation of qualia) is just an illusory, methodologically misleading and ontologically empty content, that has to be explained away. According to Edelman, it is the only means to represent the « complex differentiations » operated by the brain processes and indeed it exists in human beings as a result of evolution and cannot be dismissed as an illusory theoretical construct.

(b) This ontological difference involves a substantive epistemological difference: whereas for Dennett mechanical models are initially introduced as a metaphor but eventually, being the only promising scientific description of the data, they have to correspond to a true description of what there is—mind is a property of a complex machine—in Edelman the phenomenal content of human consciousness adequately expresses a fundamental feature of brain processes themselves. This shift can be usefully expressed in terms of models and ontology: whereas in Dennett mechanistic (computational) models of consciousness reflect a materialistic ontology, which does not leave room for any genuinely phenomenological property, in Edelman the model of the brain network reflects the phenomenological properties of consciousness itself without excluding the latter's existence. Indeed, one could even say that for Edelman consciousness itself, due to its epistemic role for human beings, is a kind of "natural model" of highly complex brain processes.

Now, going back to Descartes' legacy, one may wonder whether this distinction of consciousness from brain processes amounts to a new dualism. Regarding the distinction of consciousness from the corresponding dynamical process Edelman writes:

> « the dynamic structural origin of properties, even conscious properties, need not resemble the properties it gives rise to: an explosion does not resemble an explosive » ([14], p. 63).

Edelman's terminology is not very strict about the relation between brain and consciousness: conscious processes « emerge » from neural processes, the latter « entail » of « give rise to » conscious properties by the « phenomenal transformation » that results in qualia; qualia « reflect » neural differentiations. The view behind Edelman's theory is that consciousness is a process, whose structural properties can be traced back to structural properties of its material substratum. Since these words immediately evoke emergentism and epiphenomenalism, one is tempted to ask which ontological framework best fits the theory.

[13] See Dennett [9], pp. 84–95 and Edelman's account on his own 'Darwin' robots in Edelman [15], pp. 125–141.

5 Phenomenology and Nature: Edelman's Theory and the Problems of Contemporary Science of Consciousness

Edelman's theory of consciousness, with its deep intertwining of mechanistic models and phenomenology, presents an interesting case for contemporary science of consciousness and philosophy of mind. Though formulated as a naturalistic completion of Darwin's program, it is coherent with some anti-naturalistic claims made in contemporary phenomenological approaches. This suggests a number of methodological and ontological remarks on contemporary research.

First, Edelman's theory asserts the heuristic primacy of consciousness over mechanical explanation: the phenomenological characterization of consciousness by means of introspection is a preliminary stage of modeling and empirical research. The same conclusion has been supported in the phenomenological tradition and receives a growing attention among both philosophers and neuroscientists.[14] But agreement with this simple observation does not imply any "phenomenological" turn. The point, here, is simply that one must focus the explanandum before providing the explanation. This is even true of Dennett's theory, which aims at explaining away all the phenomenological contents, and in order to do so starts with a third-person description of subjective experience, which Dennett calls "heterophenomenology": phenomenology, therefore, must be on stage in order to be criticized. In this sense, Edelman's theory clearly shows that there is no substantial methodological contrast between naturalistic and phenomenological approaches.

Compared to the phenomenological perspective, on the other hand, the evolutionary background of Edelman's ideas allows of a different way to bridge the gap between natural science and humanities. The metaphorical and creative character of human thought adequately reflects the multidimensional system of neural activity, whose interaction with the environment is subject to the mechanism of « recognition of configuration » : insofar the description of neural networks is able to catch the overall features of experience. Nonetheless, with the development of language and culture, the properties of the « second nature » produce an autonomous domain of sense, where scientific hypotheses themselves—including neural models—are constructed, and which cannot be in itself ever subject to neurological description:

> « Although it is true that a scientific description of the world hews more closely to the structure of that world than do our daily impressions, our account of how the brain works suggests that scientific hypotheses themselves emerge from ambiguous (and occasionally irreducible) properties that give rise to pattern recognition. The brain structures and

[14] An agreement on this point was already reached by Jean-Pierre Changeux and Paul Ricoeur in their dialogue on the neurology of consciousness [6]. For a first introduction on phenomenological methods see Gallagher/Zahavi [17].

> dynamics leading to such properties are scientifically describable, even if the properties themselves cannot be fully reduced » ([15], p. 146).

This leads to the more complicated issue of ontology. By stressing the existence of qualia in a naturalistic framework Edelman provides an internal critique of reductionism, that apparently agrees with some points made by phenomenologists about the impossibility of modern science to catch the « world of life ». Indeed Edelman's theory has been appreciated by Searle, which is one of the most strenuous supporter of the impossibility to reduce phenomenology to material properties. Searle agrees with Edelman's statement that to assert the neural origin of conscious properties does not imply the latter's « reduction » to neural structures and elaborating on this point he presents an argument against Dennett's denial of qualia.[15] Nonetheless—striking as these analogies may appear—Edelman's theory presents a slightly different perspective.

The difference with Searle is signaled by a basic disagreement: Edelman considers his own model as a satisfactory scientific sample of a theory of consciousness, whereas Searle, as we have seen, considers it unsatisfactory and unable to solve the « mystery » of consciousness. But in order to understand the originality of Edelman's perspective it is useful to consider it in the context of a crucial problem of contemporary philosophy of neuroscience: that is, how to derive consciousness from a multiplicity of physical elements. Dennett himself has traced this problem back to Leibniz's claim that unity of perception cannot be caused by mechanical processes, claiming that Leibniz conflated the epistemic problem of giving this sort of explanation (which was not possible at Leibniz' times and—on the contrary—would be possible in contemporary A.I.) with an ontological verdict ([11], pp. 3–7). Dennett himself considers the subdivision of the « personal level » to « subpersonal » levels as a fundamental heuristic move. The idea behind Dennett's « intentional stance » —the attribution of conscious properties, such as belief and desire, to machines or to parts of the brain—is that

> « when we engineer a complex system (or reverse engineer a biological system like a person or a person's brain) we can make progress by breaking down the whole wonderful person into subpersons of sorts agentlike systems that have part of the prowess of a person, and then these homunculi can be broken down further into still simpler, less personlike agents, and so forth—a finite, not infinite, regress that bottoms out when we reach agents so stupid that they can be replaced by a machine » ([3], p. 88).

Now, without entering the details of Dennett's hypothesis, the claim made here is that we can conjecture a theory where physical parts compose a conscious whole, without properly attributing intentionality, or any other conscious property, to the parts themselves. As Dennett puts it, « we don't attribute fully fledged belief

[15] « Dennett denies the existence of the data to start with. But couldn't we disprove the existence of these data by proving that they are only illusions? No, you can't disprove the existence of conscious experiences by proving that they are only an appearance disguising the underlying reality, because *where consciousness is concerned the existence of the appearance is the reality* » ([26], 112).

(or decision or desire—or pain, heaven knows) to the brain parts […] No, we attribute an attenuated sort of belief and desire to these parts » ([3], p. 87). It is not entirely clear whether—and in which sense—any of these parts can be considered as a real intentional entity; Dennett himself considers the issue unimportant, since « the security of our intentional attributions at the highest levels does not depend on our identifying a lowest level of real intentionality ». This implies, according to Dennett, that we do not need to give up our standard concept of matter and can rely on the underlying scientific theories.

By making this claim Dennett is perfectly aware of an alternative way, which is to look for new physical (or psycho-physical) theory and, consequently, to modify our basic physical concept of matter. This bold move is made in a number of different ways in contemporary research: without going back to Eccles' « psychons », one can think of Penrose's « microtubules » hypothesis and of Chalmers' reference to a still undeveloped physical theory where consciousness would be a fundamental property of nature. Dennett considers all this as mere speculation (see e.g.: Dennett [9], pp. 36–37; Dennett [11], pp. 8–10), and yet this kind of speculation has found some support by one of the leading researchers in the neurology of consciousness.

Antonio Damasio considers a monistic ontology as the only reasonable background for a solution of the mind–body problem. In order to find a philosophical framework to this conviction Damasio has positively reconsidered monistic metaphysics of the past, drawing from ideas of Spinoza and Whitehead for his theory of the self.[16] Even though he does not accept any of these metaphysical theories as such, Damasio—breaking with his own dualistic terminology of the past—in his last book *Self comes to Mind* presents extension as an attribute of the mind and terminologically identifies 'neural patterns', 'maps' and 'images' ([8], p. 15, 64). This is a significant step, compared to the previous recognition of a « isomorphism » between images, neural patterns and objects ([7], p. 200). Now isomorphism is considered as a sign of objective identity. In a footnote Damasio cautiously questions the « traditional conceptions of matter and mental » as « unnecessarily narrow » ([8], p. 322, n. 14). Though recognizing that « the burden of proof does rest with those who find it natural for mind states to be constituted by brain activity », Damasio does not hesitate to set out his hypothesis: the « looped circuit » of signals transmitted between body and nervous system would enact a « functional fusion of body states and perceptual states » and—by going still deeper into the neuron circuit level—it would be possible to attribute a « protocognition » and « protofeeling » to single neurons, whose joint activity creates the conscious mind.[17]

[16] See e.g. Damasio [7], pp. 184–220, 308n.

[17] Damasio [8], pp. 256–258. While « protocognition » would correspond to the synchronic activity of a nested hierarchy of neurons (p. 252), « protofeeling » would depend on the inherent « sensitivity » or « irritability » of single cells, itself corresponding to the ability to detect and respond to changes inside and outside the cell membrane, that simple organisms display in order to preserve the homeostasis and protect the integrity of the living tissue (pp. 31–60; 258).

Here we find an interesting (and surprising) analogy with Dennett: the parts of the nervous system are endowed with a "protocognition" and "protofeeling". A huge difference lies of course in the completely different scientific framework of the two hypotheses: so Damasio really regards his line of inquiry about the properties of organic matter as « worth pursuing » , whereas Dennett relies on artificial network models and does not engage the speculative issue of the "consciousness" of cells.

These different speculations may be considered alternatively as a sign of advancement or disorientation, but there is anyway an evident problem that they all must face. That is, since we still do not have a fully articulated and successful theory of consciousness as grounded on subsytems (whether biological or artificial), both possibilities are logically open: to change the scientific theory or to change the basic concepts. As long as the hypotheses cannot rely on decisive empirical evidence, the conflict between Dennett's "multidraft" model (grounded on engineering and A.I. models) or Damasio's self theory (grounded on anatomical and biological data)—and, for that matter, Penrose's speculations on quantum–mechanical foundations of mind—cannot be settled in any definitive way.[18] Even though a number of plausibility claims can be made, and even though—for example—Damasio relies on some more factual evidence than Dennett, there is no way to decide which kind of hypothesis will lead towards a full-fledged theory of consciousness, as being grounded in physical parts of the organism. This doubles the uncertainty, as the methodological doubt is connected to a doubt about the ontological (or metaphysical) background of scientific description.

In this fragile and open context, lest we do not dismiss the whole contemporary research as « frustrating » because the subject is plagued with old mistakes ([26], p. xi),[19] and wait for a solution to be found by ways of physical or metaphysical

[18] Penrose's speculation involves the interpretation of some problematic features of Quantum mechanics, and therefore seems to add problems to problems. In quantum mechanics itself there is a similar (but, in a sense, reversed) epistemological problem: the standard theory includes a problematic interaction between observer and physical system, which has offered space for speculation and criticism; on the other hand, alternative theories (such as Bohmian Mechanics and Collapse models) are not supported by better empirical evidence and involve different conceptual and mathematical problems. For an overview see Pecere [24].

[19] Commenting on Damasio's last book, Searle criticizes his distinction between mind and consciousness, and denies that Damasio's book presents any advancement towards the solution of the « mystery of consciousness » [27]. I think that Searle, sticking to the « standard understanding of the causal relation between mind and brain », misses Damasio's point, which is to radically object to this standard view and elaborate a monistic ontology, where mind and neural patterns are two aspects of the same process which underlies consciousness. This does not mean that Damasio's theory of consciousness, which I cannot examine here in detail, is complete and free of argumentative problems (indeed, it is not). But I think that Searle's reasons of dissatisfaction lead too hastily to the usual conclusion, repeated in reply to a reader of the quoted review: « the way neurons produce consciousness remains mysterious » and « we may never have a solution to the mystery of consciousness ».

speculation (with the risk of granting the views of those who challenge the validity of standard scientific inquiry), Edelman's theory appears to offer a provisionary, reasonable standpoint, developed along the classic pathway of post-Newtonian methodology of natural science. He draws a parallel between his own experimentally provable correlation of conscious states (C) with neural states (C') and the proportionality set by the formula: F = mA ([14], p. 146). Trying to set up a measurement system, Edelman considers consciousness as a matter of experimental evidence and a property which comes in different degrees. It can be observed that, in the classical Newtonian framework, this means not to address consciousness as an essential property, while leaving open to successive inquiries whether it can have a further explanation. Nevertheless consciousness appears as a true property of organic matter, which still awaits a better understanding, but which—being measurable and possibly subject to a lawlike description—is not in itself a mystery. Here is Edelman's "Newtonian" reply to the charge of not having explained the "actual feeling of a quale":

> « these are the properties of the phenotype, and any phenotype that is conscious experiences its own differential qualia because those qualia are the distinctions made. It suffices to explain the bases of these distinctions—just as it suffices in physics to give an account of matter and energy, not why there is something rather than nothing » ([14], p. 146).

These claims can be made without committing to a particular ontological framework, such as epiphenomenalism (notwithstanding some evident similarities to this approach which are undeniable in Edelman's writings, such as his denial of conscious causality[20]).

We get therefore to some philosophical conclusions that all the quoted antagonists in contemporary research may share, since they do not require to completely settle neither the scientific, nor the ontological issues that we have discussed: consciousness is no mystery, although the description of its neural correlates is still in its early development (and could modify by way of this development our scientific or ontological tenets); but even when possessing such a description, we would not have dismissed the phenomenal and linguistic reality as a fundamental and irreducible feature of our experience. Thereby Edelman's theory is able to provide—better than Dennett's controversial and ontologically more committed account—a naturalistic background to contemporary research on the neural correlates of consciousness.

[20] For Edelman's own reply to the « charge » of epiphenomenalism see Edelman [14], pp. 81–85, 145. Edelman denies causal interaction between consciousness and brain processes, being consciousness « entailed » by these very processes. On the other hand, he denies that we are automata, because of the variability of consciousness as a reflection of the complex interaction of the plastic brain with the environment. Moreover he does not deny the role of secondary (language based) consciousness in long term planning.

References

1. Adam, C., Tannery, P. (eds.): Oeuvres de Descartes. Vrin, Paris (1964–1974)
2. Bennett, P., Hacker, P.M.S.: Philosophical Foundations of Neuroscience. Blackwell, Malden (2003)
3. Bennett, M., Dennett, D., Hacker, P., Searle, J.: Neuroscience and Philosophy. Columbia University Press, New York (2007)
4. Brook, A., Ross, D. (eds.): Daniel Dennett. Cambridge University Press, Cambridge (2002)
5. Changeux, J.-P.: L'homme neuronale. Fayard, Paris (1983)
6. Changeux, J.-P., Ricoeur, P.: Ce qui nous fait penser. La nature et la régle. Odile Jacob, Paris (1996)
7. Damasio, A.: Looking for Spinoza. Vintage Books, London (2004)
8. Damasio, A.: Self Comes to Mind. Constructing the Conscious Brain. Pantheon Books, New York (2010)
9. Dennett, D.: Consciousness Explained. Little, Brown & Co., Boston (1991)
10. Dennett, D.C.: With a little help from my friends. In: Ross, D., Brook, A., Thompson, D. (eds.) Dennett's Philosophy. A Comprehensive Assessment, pp. 327–388. MIT Press, Cambridge (2000)
11. Dennett, D.C.: Sweet Dreams. Philosophical Obstacles to a Science of Consciousness, MIT Press, Cambridge (2005)
12. Edelman, G.M.: Remembered Present. A Biological Theory of Consciousness. Basic Books, New York (1989)
13. Edelman, G.M.: Bright Air, Brilliant Fire: On the Matter of the Mind. Basic Books, New York (1992)
14. Edelman, G.M.: Wider than the Sky. The Phenomenal Gift of Consciousness. Yale University Press, Yale (2004)
15. Edelman, G.M.: Second Nature. Brain Science and Human Knowledge. Yale University Press, New Haven (2006)
16. Edelman, G.M., Tononi, G.: A Universe of Consciousness. How Matter Becomes Imagination. Basic Books, New York (2000)
17. Gallagher, S., Zahavi, D.: The Phenomenological Mind. Routledge, Abingdon (Oxon)/New York (2008)
18. Garber, D.: Understanding interaction. What Descartes should have told Elisabeth. In: Id., Descartes Embodied. Reading Cartesian Philosophy through Cartesian Science. Cambridge University Press, Cambridge (2001)
19. Leibniz, G.W.F.: Monadology. In: Gerhardt, C.J. (hrsg.), Die philosophischen Schriften, vol. VI, pp. 607–623. Lorentz, Leipzig (1932)
20. McLaughlin, B.P.: The Oxford Handbook of Philosophy of Mind, Oxford University Press, Oxford (2009)
21. Moravia, S.: The Enigma of the Mind. The Mind-Body Problem in Contemporary Thought. Cambridge University Press, Cambridge (1995)
22. Stenon, N.: Discours sur l'anatomie du cerveau. Robert de Ninville, Paris (1669)
23. Tononi, G.: An information integration theory of consciousness. BMC Neurosci. **5**, 42 (2004). http://www.biomedcentral.com/1471-2202/5/42
24. Pecere, P.: Fisica quantistica e realtà. Considerazioni storico-filosofiche. In: Argentieri, N., Bassi, A., Pecere, P., Meccanica quantistica rappresentazione realtà. Un dialogo tra fisica e filosofia. Bibliopolis, Napoli (2012)
25. Schneider, S.: Daniel Dennett on the nature of consciousness. In: Velmans/Schneider (eds.): The Blackwell Companion to Consciousness, Blackwell, Malden, pp. 313–324 (2007)
26. Searle, J.: The Mystery of Consciousness. The New York Book Review, New York (1997)
27. Searle, J.: The Mystery of Consciousness Continues. In: « The New York Review of Books » , June 11, 2011, http://www.nybooks.com/articles/archives/2011/jun/09/mystery-consciousness-continues/?pagination=false#fnr-2

28. Velmans, M., Schneider, S. (eds.): The Blackwell Companion to Consciousness, Blackwell, Malden (2007)
29. Voss, S.: Descartes: Heart and Soul. In: Wright, J.P., Potter, P. (eds.) Psyche and Soma. Physicians and Metaphysicians on the Mind-Body Problem from Antiquity to Enlightenment, Oxford University Press, Oxford (2000)

Information Integration in the Brain: Modeling a Putative Function of the Astroglial Network

Alfredo Pereira Jr and Fábio Augusto Furlan

Abstract Astrocytes receive somatic signals carried by blood flow and cerebrospinal fluid, as well as sensory and cognitive information carried by neuronal assemblies. Their position, hub-like structure and intrinsic processing capabilities suggest that they integrate spatially distributed information. Oscillatory synchrony and constructive wave interference have been proposed to constitute mechanisms of neuro-astroglial interaction. We further claim that the astrocytic network appreciates information contents processed by neurons, modulating neuronal activity (and then the behavior) according to valences attributed to them. Here we report our modeling of mental functions of astrocytes and discuss empirical results that could confirm or desconfirm the model.

1 Introduction

In perceptual processes of human individuals and other species, signals carrying information about a stimulus are transmitted through multiple processing lines to populations of receptive neurons and thalamo-cortical circuits, leading to the formation of a spatial ensemble of local field potentials (LFPs). Our proposed scientific model [20–25] addresses the problem of how the brain (more specifically, the astroglial network interacting with distributed neuronal assemblies) integrates patterns embodied in local fields to (re)construct the stimulus in a conscious episode.

A. Pereira Jr (✉)
Institute of Biosciences, São Paulo State University (UNESP), Botucatu, Brazil
e-mail: apj@ibb.unesp.br

F. A. Furlan
University of Marilia (UNIMAR)—School of Medicine, Marília, Brazil
e-mail: fabioaugustofurlan@yahoo.com.br

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_10, © Springer-Verlag Berlin Heidelberg 2014

In the mammalian brain, together with the development of neocortex and thalamo-cortical-striatal circuits, specialized neuronal circuits process endogenous and exogenous information in a distributed way. At the micro level, proteins located at the membrane of neurons—the receptors—receive information from the body and external world, control fluxes of ions and produce coherent electric fields that putatively reproduce aspects of the stimuli and support cognitive processes.

In the dominant neuroscience paradigm, these electric fields have been partially and indirectly measured in several ways (single cell recording, EEG, fMRI, optical imaging), and experimentally correlated to the content of reports made by the individual about his/her experiences or to the behavior of animals. The same neurotransmitters that generate the electric fields can elicit—by a variety of pathways—calcium waves in astrocytes, which endfeet wrap the synapses [21]. We have proposed that these waves integrate neuronal spatially distributed information and instantiate feelings, leading to the attribution of valences to the contents processed by neurons, putatively supporting conscious episodes experienced by the living individual. Astrocytes do not have direct access to the generation of behavior or to the mechanisms of memory formation, but can influence them by means of modulatory actions on neurons in the context of the tripartite synapse.

The development of our model is based on the current state of the art of research on cognitive and affective functions of the astrocyte network. Regarding the relationship between astrocyte activity and human consciousness, besides the fundamental discoveries of Oberheim et al. [18, 19] on types of astrocytes unique to our species (reviewed in [21]), it is worth noting a recent discovery of Brazilian neuroanatomists (see [14]), pointing towards the existence of a greater amount of glial cells in brain regions correlated with conscious activity, while a higher proportion of neurons is found in the cerebellum, which has little or no contribution to conscious processing.

2 Basic Mechanisms of Neuro-Astroglial Interactions

Glutamate (Glu) is the main excitatory transmitter in the brain, being largely present in cortico-cortical networks and operating both on excitatory (as pyramidal cortical) and inhibitory neurons (as GABAergic interneurons). Glu operates as an information carrier to thalamocortical and cortico-cortical synapses, a role that is crucial for the understanding of perceptual processing in the brain. Activation of three kinds of Glu receptors (NMDA, AMPA and metabotropic) converges to each neuron's dendritic spine, where calcium ions (Ca^{2+}) entering through NMDA control Calmodulin (CaM) and Calmodulin-Dependent Protein Kinase II (CaM-KII) regulatory mechanisms.

Glutamatergic synapses are mostly tripartite. The Glu released by the presynaptic neuron reaches both the postsynaptic neuron and the astrocyte. By binding to the NMDA and AMPA receptors of the postsynaptic neuron, Glu induces membrane depolarization, and the opening of NMDA channels to fast calcium currents. This excitation decays in around 150 ms [10], but can be sustained by gliotransmission (Glu released by the astrocyte, binding to extrasynaptic NMDA receptors), producing slow calcium currents through NMDA receptors. The second input of calcium may activate the path of calmodulin and its kinase (CaMKII), which phosphorylates or dephosphorylates AMPA receptors and thus induces synaptic potentiation or depression.

While the individual is awake, the astrocytic network is pre-activated by purinergic and cholinergic mechanisms, facilitating the generation of calcium waves at the time when Glu transmission occurs. In the astrocyte, the induction of local calcium waves by means of excitatory transmission (by another transmitter or Glu) may generate intercellular, global waves that broadcast information to many other parts of the brain [21].

We have proposed two mechanisms of formation of global calcium waves in the astrocyte network [21]. The *domino effect* explains signal propagation in astrocytic calcium waves by transfering the vibrational energy from ion to ion, and signal amplification by ATP both in gap junctions and in extracellular mediums, without requiring displacement of the ions between the microdomains. The *carousel effect* explains how synchronized neuronal activity induces a large-scale calcium wave in astrocytic network, which in the next moment modulates neuronal activity. The “domino” and “carousel” effects conjointly explain how patterns embodied in activity of synchronized neuronal networks can be readily transferred to calcium waves in the astroglial network.

After sensory patterns are embodied in calcium waveforms, their integration takes more then 100 ms to occur. A different case is conscious reasoning, like perceiving a grammatical mistake in a visually presented sentence. In this case, the integration process in the astroglial network is more complex and takes additional hundreds of milliseconds, corresponding to the time course of the respective evoked potentials. A parsimonious explanation of our capacity of operating both unconsciously at millisecond times and consciously at the scale of seconds is a combination of neuronal and astroglial processing in superposed time scales. The neuro-astroglial interaction mechanisms operate with these two timescales, one neuronal (at the range of milliseconds) and other astroglial (in the range of seconds to minutes). Event-Related Potentials (ERPs) correlated with conscious events take from 100 to 1000 ms to occur. If their generation depended only on neuronal transmission through cortico-cortical axons, ERPs would take only 50–100 ms. The astroglial timing also corresponds to the Slow Cortical Potential [11] described for BOLD fMRI results, considering both positive correlations (between percepts and fMRI activations) and default networks. This correspondence is not surprising, since astrocytes exert the control of blood flow [26].

What is the role of calcium waves in conscious processing? Pereira [25] advanced the hypothesis that calcium ions trapped in the astroglial endoplasmatic reticulum and intracellular microdomains interact and form local correlations. These microdomains are connected to each other by means of gap junctions and also communicate sequentially by means of extracellular ATP signaling. When a synchronized population of neurons activates—by means of Glu, other neurotransmitters or electromagnetically—the astroglial network, the previous correlations possibly play the role of tying ions in the same vibrational mode, operating as antennas that broadcast the pattern received from local neuronal fields to the whole astroglial network, forming a kind of hologram that embodies integrated information and results in a feeling of the experienced episode, which feeds back on neuronal networks [25].

3 The Relevance of Neuro-Astroglial Interactions for Conscious Processes

The processing of sensory information patterns and their combination into temporal conscious episodes is a complex process with several phases. In humans and possibly other animal species, the integration of dynamic ionic patterns processed by neuronal assemblies into ‘gestalts’ (conscious episodes) requires the broadcasting and integration of the patterns carried by feedforward/feedback neuronal circuits. The pattern of a stimulus (a three-dimensional object and/or invariants in a dynamical interaction process) is transduced by sensory receptors (e.g. in the cells of the retina and/or the olfactory bulb) to electric patterns. Signals from peripheral sensors (or central sensors, in the case of the retina) to the central nervous system (CNS) are carried by nerves, using a population frequency code. In the CNS, perceptual processing begins with single-neuron feature detection, filtering of salient features, local broadcasting and activation of specialized circuits. This initial sensory processing phase leads to global feed-forward broadcasting by means of axonal vertical (cortico-thalamic) and horizontal (cortico-cortical) connections, and feedback from associative to sensory areas. The feedback has an adaptive function of making the sensory pattern salient in the respective context.

The form of the stimulus is decomposed into an ensemble of signals, each one presumably reproducing an aspect of the object and/or process being perceived. In the CNS, these patterns are embodied in receptive fields, by means of an activation of the neuronal dendritic graded potential. These potentials are generated mostly by ligand-gated ion channels (e.g., AMPA glutamatergic receptors) that control the movements of ions through neuronal membranes. At this stage, the sensory message about aspects of the stimulus elicits the formation of an ensemble of Local Field Potentials (LFP) located at several cortical areas. Considering the spatial distribution of LFPs and their effects on the astroglial network, the latter is in an excellent condition to integrate and produce feelings about the content of the information. A sketch of the whole proposed process leading to the formation of a conscious percept is illustrated in Fig. 1.

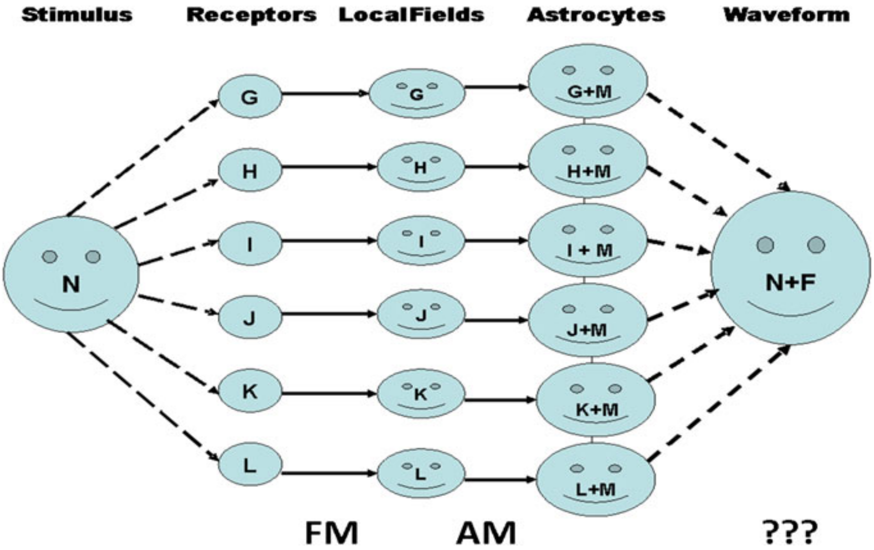


Fig. 1 *Overview of the Process of Cognitive Computation Leading to a Conscious Percept.* A given stimulus has N properties (including *G*, *H*, *I*, *J*, *K*, *L*), separately detected and processed by different sensory circuits in the brain, leading to the formation of a spatial ensemble of LFPs, each one reproducing one aspect of the stimulus (*G*, *H*, *I*, *J*, *K*, *L*). Upon transmission of these patterns to astrocytes, these cells attach to the information content a subconscious meaning (*M*), relative to the history of the living individual. Psychophysically, these meanings correspond to small calcium waves located in microdomains of astrocytes. The integration of these small waves results in a global wave that embodies the conscious percept, thus reproducing the *N* processed properties and adding to this content a feeling (*F*). The information is frequency modulated (*FM*) in neuronal axonal transmissions and also amplitude modulated (*AM*) both in neuronal dendritic potentials and astroglial calcium waves. The final waveform that corresponds to the conscious feeling of the stimulus possibly has frequency, amplitude and phase dynamical structures, and possibly involves other mechanisms that are still not well understood, such as quantum entanglement (see Pereira [25])

4 Implications of the Model and the Need of Corroborations

Our model is based on tripartite synapses, covering their molecular, systemic and functional aspects, and focusing on cognitive and affective functions of astrocytes in the context of their interactions with neurons. The functions of integration and appreciation of information are considered to be closely related, because the integration mode (for example, performing a figure-ground distinction) already contains an appraisal (for example, by assigning positive or negative valence to certain aspects of the figure). Our published model describes how the astroglial network participates in processes of integration and appreciation of information patterns processed in a distributed manner in the thalamocortical system, while interacting with somatic processes by means of signaling via blood flow and cerebral fluid, and uses the results to modulate neuronal network, thus participating in processes of perception, attention, learning, formation of semantic and episodic memory, emotion, consciousness and behavior.

While in the "neuron doctrine" (formulated by Ramon y Cajal) neurons were considered the functional units of the mind, in the new emerging model of neuro-astroglial interactions the tripartite synapse becomes the functional unit, constituting the basis for psycho-physiological processes. As an important consequence for research in brain science, it is suggested that mental functions must be experimentally correlated with neuro-astroglial processes, and not just with neuronal activity.

In the published model, neural network are mostly responsible for cognitive processes (such as the formation of representations and logical operations in processes of perception, attention, action, learning and memory formation), whereas the astrocytic network performs an appreciation of the information carried by the neural network (e.g., in terms of "pleasant" or "unpleasant", "attractive" or "disgusting"). The integration-and-appreciation process is achieved by means of intercellular, global calcium waves. Based on the appreciation, the astrocytic network modulates neuronal activity, reinforcing the patterns of information that have positive valence for the individual, and weakening those with negative valence.

Recently, we made a synthesis or our model with the psycho-physiological model developed by Claudia Carrara-Augustenborg, entitled "Endogenous Feedback Network [2, 6]." We conceive that this endogenous feedback involves the interaction of neuronal and astroglial networks, the first responsible for the processing of information and the constructing of representations, while the second would be responsible for the evaluation of the content of the information that is processed [25].

There has been considerable debate in the scientific community regarding the existence and possible roles of calcium waves 'in vivo' (for an introduction to the subject, see [27]). The hypothesis was recently confirmed by three publications [13, 17, 30], but the consequences of these findings have not yet been absorbed by a large majority of neuroscientists. The news point towards a paradigm shift in brain science (as predicted by Bullock et al. [5]; Douglas Fields, 2009), exceeding the limits of the "neuron doctrine" and moving towards the development of models that consider neuron-astrocyte interactions (both at the molecular level—tripartite synapse—as at a systemic level) as the domain to model cognitive and affective processes experienced by biological individuals in the context of interaction with their physical, biological, social and cultural environment.

Notwithstanding such corroboration, a full validation of the model requires further evidence to the effect that the astrocytic network would be effectively responsible for evaluating the information processed by neurons. Experiments to test this hypothesis have great technical difficulty, since it would be necessary to decouple measurements of neuronal and astroglial activities. It is known that both BOLD fMRI (see [26]) as scalp electroencephalogram [3] measure these activities together.

Although astrocytes maintain their membrane in hyperpolarized states, not generating action potentials, they exhibit excitatory activity (waves of calcium ions), which also produce electrical currents and their orthogonal electromagnetic fields detected by electrodes placed on the scalp. From this physiological activity, astrocytes induce the release of gliotransmissores, molecules and ions, modulate neuronal activity and exert vascular control [10], determining the magnitude of the hemodynamic response measured by fMRI.

However, there is still no method that allows separating the components of the fMRI and/or EEG generated by each cell type, neuronal or astroglial (as in general connectivity measurements [28]). Measurement of slow waves—below 1 Hz—generated by astrocytic activity would require special techniques and equipment, such as subdural electrodes used for detecting certain types of seizures. These methods, used in clinical medicine, present major difficulties for use in psychological research.

Imaging of astrocytic calcium waves 'in vitro' by optical means (for example, see http://www.youtube.com/watch?v=SJGI3dMvBlI), raised the interest in the phenomenon, leading experts to wonder about the functionality of the waves (for example, see [1]). In order to better understand the function of these waves, it is necessary to conduct studies 'in vivo'. Years ago, researchers [12] were able to generate images of waves using fluorescence microscopy ("two-photon microscopy", a technique in which a beam of light is projected on brain tissue, allowing the microscope to capture reflectance patterns generated by astrocytic calcium waves).

Movies with better resolution have been made using a sophisticated technology. Fluorescent fish genes are inserted into the DNA of transgenic mice, in the sectors responsible for the expression of the calcium ion receptor proteins. The physiological activity of the brain of these animals display increased brightness when calcium ions bind to such receptors. Through a 'window' opened on the skull of transgenic mice, it is possible to view (using a microscope) the calcium waves.

The use of transgenic animals with a deficit of functional proteins in astrocytes has led to controversial interpretation of results. Animals with deficient expression of connexins have major behavioral disabilities (James Robertson, personal communication), and Han et al. [9] demonstrated that in transgenic mice lacking cannabinoids receptors in astrocytes, a working memory deficit promoted by marijuana was abolished. Nevertheless, animals with damage in the inositol triphosphate signal transduction pathway apparently do not have notable losses of sensory, cognitive or motor functions (see discussion in [27]). A possible interpretation of this result is that such animals would compensate the shortage of glutamatergic transmission and signal transduction by means of other transmitters such as cholinergic and purinergic ones. In this sense, it is important to emphasize that two confirmations of the existence of astrocytic calcium waves 'in vivo' concerns cholinergic transmission [17, 30]. Some authors attribute a central role to purinergic transmission [32], which could hardly be blocked in transgenic animals without disturbing their vital processes.

5 Perspectives for Future Research

Given the technological limitations in the measurement and imaging of astrocytic calcium waves, we are developing three indirect experimental strategies, which may contribute to a better understanding of its role in psychological processes:

(a) Project 1—Production of a prototype of one artificial astrocyte, which reproduces its basic physiological phenomenon: the formation, propagation and interaction of calcium waves (as reported in Pereira [25]), also taking into consideration the water solution of these ions that occur 'in vivo' (see [16]), but without the proteins that control the ionic solution;

(b) Project 2—Experiment of cortical stimulation with weak electrical currents (following the recommendations of Tadini et al. [29]) containing musical forms, and evaluation of its subliminal and/or supraliminal effects. The use of music is due to its ability to activate the functions of integration and appreciation, which—according to our hypothesis—would be carried by the astrocytic network.

(c) Project 3—Behavioral study of the formation of semantic memory (in continuation of [15] and [4]), considering the role of astrocytic network in strengthening or weakening of information patterns, in accordance with the model of De Pittà et al. [7].

6 Concluding Remarks

Recent results in brain physiology and pharmacology research (specially [31]) suggest that at least a part of conscious processing is dependent on astroglial calcium waves. In a Monist philosophical perspective, this finding bypasses common worries about the Hard Problem (i.e., how to explain first-person conscious experience in a physical world?), pointing towards the concept of consciousness as a composite process of cognitive representations instantiated in neurons and feelings instantiated in the astroglial network. The view that emerges is that some physical activities in our body are the objective side of our subjective first-person conscious experiences. A practical consequence is that our knowledge about these activities gives us powers to better understand and treat the subjective side. Astrocytes have also been implicated in the aetiology and therapy of most psychiatric and neurological problems, for instance in Alzheimer's [8].

Our model takes into account the dynamics of conscious processing by the living individual. In the awakening process, the astroglial network is primed by cholinergic signals, becoming ready to respond to neuronal input with the generation of small and large calcium waves. As the astrocyte is in contact with blood and cerebrospinal fluid (while neurons are not), it receives the signals that come in the flow, forming a continuously updated reference of the state of the body in the world. The latter signals generate only small calcium waves; they are 'spontaneous' in the sense of being not induced by neurons, but by energy from blood, signals from blood and cerebrospinal fluid, and endogenous processes in the astroglial network.

Astrocytes are not connected to sensory transducers or muscle and endocrine effectors. All sensation and perception, as well as all actions in the world begin with neurons; the results of neuronal processes (information patterns embodied in local electromagnetic fields) reach the astrocytes and induce the larger waves (see the 'domino' and 'carousel' effects in Pereira and Furlan [21]). The larger waves in the astrocyte network—that we model to correspond to conscious activity—require a coordinating action from neurons, by means of glutamatergic and purinergic transmission:

(a) local EM fields are formed by active neuronal assemblies;
(b) neuronal large-scale synchronization occurs in theta to gamma frequencies (not delta; synchronization in the slowest frequencies imply unconsciousness);
(c) there are chemical and ephaptic (magnetic) transmissions of information from the neuronal local fields to astroglial waves; and
(d) there is an interference of the smaller waves, leading to the formation of the larger ones in the astroglial network.

We claim that neurons do the cognitive job, generating (unconscious or subconscious) representations of the body and world, while astrocytes, which are in closer contact with the whole body (and then evaluate the information conveyed by the neurons in terms of their consequences for the survival and well-being of the body) instantiate feelings essential to conscious activity (conceptualized as "the feeling of what happens", as in the title of Damasio's book) and then modulates neuronal activity according to the valences attributed to the patterns. This psychophysical model can help philosophers of the mind to develop more realistic conjectures about brain endogenous cognitive and affective processes, the resulting conscious experiences, and the relation of consciousness with the (rest of the) world, in terms of perception and action processes.

References

1. Agulhon, C., Petravicz, J., McMullen, A.B., Sweger, E.J., Minton, S.K., Taves, S.R., Casper, K.B., Fiacco, T.A., McCarthy, K.D.: What is the role of astrocyte calcium in neurophysiology? Neuron **59**, 932–946 (2008)
2. Almada, L., Pereira, Jr., A, Carrara-Augustenborg, C.: (forthcoming) What Affective Neuroscience Means for a Science of Consciousness. Mens Sana Monogr. **11**(1), 253–273 (2013). doi:10.4103/0973-1229.100409
3. Banaclocha, M.A.M.: Neuromagnetic dialogue between neuronal minicol—umns and astroglial network: a new approach for memory and cerebral computation. Brain Res. Bull. **73**, 21–27 (2007)

4. Barros, R.F., Santos, R.P., Furlan, F.A., Camilo, L.A., Pereira Jr., A.: Efeitos de relevância versus repetição de estímulo linguístico na indução da memória declarativa. Neurociências **7**(1), 6–19 (2011)
5. Bullock, T.H., Bennett, M.V., Johnston, D., Josephson, R., Marder, E., Fields, R.D.: The neuron doctrine, redux. Science **310**(5749), 791–793 (2005)
6. Carrara-Augustenborg, C., Pereira Jr., A.: Brain endogenous feedback and degrees of consciousness. In: Cavanna, A.E., Nani, A. (eds.) Consciousness: States, Mechanisms and Disorders. Nova Science Publishers Inc, New York (2012)
7. De Pittà, M., Volman, V., Berry, H., Ben-Jacob, E.: A tale of two stories: astrocyte regulation of synaptic depression and facilitation. PLoS Comput. Biol. **7**(12), e1002293 (2011)
8. Furman, J.L., Sama, D.M., Gant, J.C., Beckett, T.L., Murphy, M.P., Bachstetter, A.D., Van Eldik, L.J., Norris, C.M.: Targeting astrocytes ameliorates neurologic changes in a mouse model of Alzheimer's disease. J. Neurosci. **32**(46), 16129–16140 (2012)
9. Han, J., Kesner, P., Metna-Laurent, M., Duan, T., Xu, L., Georges, F., Koehl, M., Abrous, D.N., Mendizabal-Zubiaga, J., Grandes, P., Liu, Q., Bai, G., Wang, W., Xiong, L., Ren, W., Marsicano, G., Zhang, X.: Acute cannabinoids impair working memory through astroglial CB_1 receptor modulation of hippocampal LTD. Cell **148**, 1039–1050 (2012)
10. Haydon, P.G., Carmignoto, G.: Astrocyte control of synaptic transmission and neurovascular coupling. Physiol. Rev. **86**, 1009–1031 (2006)
11. He, B.J., Raichle, M.E.: The fMRI signal, slow cortical potential and consciousness. Trends Cogn. Sci. **13**, 302–309 (2009)
12. Hirase, H., Qian, L., Barthó, P., Buzsáki, G.: Calcium dynamics of cortical astrocytic networks in vivo. PLoS Biol. **2**(4), E96 (2004)
13. Kuga, N., Sasaki, T., Takahara, Y., Matsuki, N., Ikegaya, Y.: Large-scale calcium waves traveling through astrocytic networks in vivo. J. Neurosci. **31**(7), 2607–2614 (2011)
14. Lent, R., Azevedo, F.A., Andrade-Moraes, C.H., Pinto, A.V.: How many neurons do you have? Some dogmas of quantitative neuroscience under revision. Eur. J. Neurosci. **35**(1), 1–9 (2012)
15. Marques, J.F., Barros, R.F., Santos, R.P., Pereira Jr, A.: Estratégias de somação temporal e espacial na formação da memória declarativa de curto prazo. Scripta (PUCMG) **14**, 43–56 (2010)
16. Mentré, P.: Water in the orchestration of the cell machinery. Some misunderstandings: a short review. J. Biol. Phys. **38**, 13–26 (2012)
17. Navarrete, M., Perea, G., de Sevilla, D.F., Gómez-Gonzalo, M., Núñez, A., et al.: Astrocytes mediate in vivo cholinergic-induced synaptic plasticity. PLoS Biol. **10**(2), 1001259 (2012)
18. Oberheim, N.A., Wang, X., Goldman, S.A., Nedergaard, M.: Astrocytic complexity distinguishes the human brain. Trends Neurosci. **29**, 547–553 (2006)
19. Oberheim, N.A., Takano, T., Han, X., He, W., Lin, J.H.C., Wang, F., Xu, Q., Wyatt, J.D., Pilcher, W., Ojemann, J., Ransom, B.R., Goldman, S.A., Nedergaard, M.: Uniquely hominid features of adult human astrocytes. J. Neurosci. **29**, 3276–3287 (2009)
20. Pereira Jr, A., Furlan, F.A.: On the role of synchrony for neuron-astrocyte interactions and perceptual conscious processing. J. Biol. Phys. **35**(4), 465–481 (2009)
21. Pereira Jr, A., Furlan, F.A.: Astrocytes and human cognition: modeling information integration and modulation of neuronal activity. Prog. Neurobiol. **92**, 405–420 (2010)
22. Pereira Jr, A., Almada, L.F.: Conceptual spaces and consciousness: integrating cognitive and affective processes. Int. J. Mach. Consci. **3**(1), 127–143 (2011)
23. Pereira Jr, A., Furlan, F.A., Pereira, M.A.O.: Recent advances in brain physiology and cognitive processing. Mens Sana Monogr. **9**, 183–192 (2011)
24. Pereira Jr, A., Furlan, F.A.: Analog modeling of human cognitive functions with tripartite synapses. Stud. Comput. Intell. **314**, 623–635 (2011)
25. Pereira Jr, A.: Perceptual information integration: hypothetical role of astrocytes. Cogn. Comput. **4**(1), 51–62 (2012)
26. Schummers, J., Yu, H., Sur, M.: Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. Science **320**, 1638–1643 (2008)

27. Smith, K.: Settling the great glia debate. Nature **468**, 160–162 (2010)
28. Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A.G., Bruno, M.A., Mariotti, M., Boveroux, P., Tononi, G., Laureys, S., Massimini, M.: Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. Brain **135**(4), 1308–1320 (2012)
29. Tadini, L., El-Nazer, R., Brunoni, A.R., Williams, J., Carvas, M., Boggio, P., Priori, A., Pascual-Leone, A., Fregni, F.: Cognitive, mood, and electroencephalographic effects of noninvasive cortical stimulation with weak electrical currents. J. ECT. **27**(2), 134–140 (2010)
30. Takata, N., Mishima, T., Hisatsune, C., Nagai, T., Ebisui, E., Mikoshiba, K., Hirase, H.: Astrocyte calcium signaling transforms cholinergic modulation to cortical plasticity in vivo. J. Neurosci. **31**(49), 18155–18165 (2011)
31. Thrane, A.S., Thrane, V.R., Zeppenfeld, D., Lou, N., Xu, Q., Nagelhus, E.A., Nedergaard, M.: General anesthesia selectively disrupts astrocyte calcium signaling in the awake mouse cortex. Proc. Natl. Acad. Sci. USA [Epub ahead of print] (2012)
32. Verderio, C., Matteoli, M.: ATP in neuron-glia bidirectional signalling. Brain Res. Rev. **66**(1–2), 106–114 (2011)

Part II
Abduction, Problem Solving and Practical Reasoning

Understanding Abduction

Inference, Perception, and Instinct

Lorenzo Magnani

> Unless man has a natural bent in accordance with nature's, he has no chance of understanding nature at all.
> Charles Sanders Peirce, *A Neglected Argument for the Reality of God*, 1908.

Abstract The status of abduction is still controversial. When dealing with abductive reasoning misinterpretations and equivocations are common. What did Peirce mean when he considered abduction both a kind of inference and a kind of instinct or when he considered perception a kind of abduction? Does abduction involve only the generation of hypotheses or their evaluation too? Are the criteria for the best explanation in abductive reasoning epistemic, or pragmatic, or both? Does abduction preserve ignorance or extend truth or both? To study some of these conundrums and to better understand the concept of abduction, which Hintikka [20] classified the "fundamental problem of contemporary epistemology", I think that an interdisciplinary effort is needed, at the same time fecundated by a wide philosophical analysis. To this aim I will take advantage of some reflections upon Peirce's philosophy of abduction that I consider central to highlight the complexity of the concept, too often seen in the partial perspective of limited (even if tremendously epistemologically useful) formal and computational models. I will ponder over some seminal Peircean philosophical considerations concerning the entanglement of abduction, perception, inference, and instinct, which I consider are still important to current cognitive research. Peircean analysis helps us to better grasp how model-based, sentential and manipulative aspects of abduction—I have introduced in my book *Abductive Cognition* [38]—have to be seen as intertwined, and indispensable for building an acceptable integrated model of abduction. Even if speculative, Peircean philosophical results on abduction certainly anticipate various tenets of recent cognitive research, as I will remark.

L. Magnani (✉)
Department of Humanities, Philosophy Section, and Computational Philosophy Laboratory,
University of Pavia, Pavia, Italy
e-mail: lmagnani@unipv.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_11, © Springer-Verlag Berlin Heidelberg 2014

1 Iconicity and Logicality in Abductive Cognition

1.1 Perception Versus Inference?

We should remember, as Peirce noted, that abduction plays a role even in relatively simple visual phenomena. *Visual abduction*,[1] a special form of non verbal abduction—a kind of model-based cognition—occurs when hypotheses are instantly derived from a stored series of previous similar experiences. It covers a mental procedure that falls into the category called “perception”. Peirce considers *perception* a fast and uncontrolled knowledge-production process. Perception is a kind of vehicle for the instantaneous retrieval of knowledge that was previously assembled in our mind through inferential processes. Keeping in mind Peirce’s famous syllogistic framework for abduction (as a form of fallacy of the affirming the consequent) we can say that, in the case of perception we face with a situation in which: “[...] a fully accepted, simple, and interesting inference tends to obliterate all recognition of the uninteresting and complex premises from which it was derived” [49, 7.37]. We can add that many visual stimuli—that can be considered the “premises” of the involved abduction—are ambiguous, yet people are adept at imposing order on them: “We readily form such hypotheses as that an obscurely seen face belongs to a friend of ours, because we can thereby explain what has been observed” [67, p. 53]. This kind of image-based hypothesis formation can be considered as a form of *visual abduction*. Hence, perception is abductive in itself: “Abductive inference shades into perceptual judgment without any sharp line of demarcation between them” [54, p. 224]. Visual abduction plays an important cognitive role in both everyday reasoning and science, where it is well known it can provide epistemically substantial shortcuts to dramatic new discoveries.

If perceptions are abductions they are basically withdrawable, just like the scientific hypotheses abductively found. In this perspective perceptions can be seen as “hypotheses” about data we can accept (usually this happens spontaneously) or carefully evaluate. Moreover, the fact they can be considered, as we will see in the following subsection, *inferences*, in the Peircean sense, and so withdrawable, does not mean they are controlled (deliberate), like in the case of explicit inferences, for example in logic and other types of rational or fully conscious human reasoning. Perception involves semiosis and is abductive, and it is able to correct itself when it falls into error, and consequently it can be censured. However, we have to carefully analyze the proper character of this kind of controllability, following Peirce’s considerations on the so-called “perceptual judgment” (“The seven systems of metaphysics”, 1903):

[1] I have introduced visual abduction in [34, 35].

> Where then in the process of cognition does the possibility of controlling it begin? Certainly not before the *percept* is formed. Even after the percept is formed there is an operation, which seems to me to be quite uncontrollable. It is that of judging what it is that the person perceives. A judgment is an act of formation of a mental proposition combined with an adoption of it or act of assent to it. A percept on the other hand is an image or moving picture or other exhibition. [...] I do not see that it is possible to exercise any control over that operation or to subject it to criticism. If we can criticize it at all, as far as I can see, that criticism would be limited to performing it again and seeing whether with closer attention we get the same result. But when we so perform it again, paying now closer attention, the percept is presumably not such it was before. I do not see what other means we have of knowing whether it is the same as it was before or not, except by comparing the former perceptual judgment to the later one. I would utterly distrust any other method of ascertaining what the character of the percept was. Consequently, until I am better advised, I shall consider the *perceptual judgment* to be utterly beyond control [54, II, p. 191].

In summary, judgments in perception are fallible but indubitable abductions—we are not in any condition to psychologically conceive that they are false, as they are unconscious habits of inference.

Nevertheless, percept and perceptual judgment are not unrelated to abduction because they are not entirely free

> [...] from any character that is proper to interpretations [...]. The fact is that it is not necessary to go beyond ordinary observations of common life to find a variety of widely different ways in which perception is interpretative. The whole series of hypnotic phenomena, of which so many fall within the realm of ordinary everyday observation,—such as waking up at the hour we wish to wake much nearer than our waking selves could guess it, —involve the fact that we perceive what we are adjusted for interpreting though it be far less perceptible that any express effort could enable us to perceive [...]. It is a marvel to me that the clock in my study strikes every half an hour in the most audible manner, and yet I never hear it [...]. Some politicians think it is a clever thing to convey an idea which they carefully abstain from stating in words. The result is that a reporter is ready to swear quite sincerely that a politician said something to him which the politician was most careful not to say. It is plainly nothing but the extremest case of Abductive Judgment [54, II, p. 229].

1.1.1 Icons, Perceptions, and Model-Based Abduction

The fact that perception functions as a kind of "abstractive observation" [54, II, p. 206], so that "perceptual judgments contain general elements" [54, II, p. 227] relates it to the expressive power of icons. It is analogous to what is occurring in mathematics when the reasoner "sees"—through the manipulations and constructions on an external single diagram (icon)—that some properties are not merely single but of a general nature: perception functions as "an abstractive observation". Indeed Peirce was clearly aware, speaking of the model-based aspects of deductive reasoning, that there is an "experimenting upon this image [for example 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" [49, 3.363]. Peirce eloquently concludes that it is "[...] a very extraordinary nature of Diagrams that they show—as literally as Percept shows the Perceptual Judgment to be true,—that a consequence does follow, and more marvelously yet, that it would follow under all varieties of circumstances accompanying the premises" [51, pp. 317–318].[2]

These Peircean considerations also echo the Kantian ones concerning geometry. Immanuel Kant was clearly aware of the interplay between internal and external models—exemplified in the case of a formal science like mathematics—as an example of genuine knowledge production (and, occasionally, of discovery). In his 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" [24, 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", and exploit "external" representations: to solve the classical geometrical problem of the sum of the internal angles of a triangle, the agent for example "[...] 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" ([24, A716-B744, pp. 578–579], emphasis added).

1.1.2 Emotions as Abductions

One more example that supports this interpretative nature of perception is given by the fact that the perception of tone arises from the activity of the mind only after having noted the rapidity of the vibrations of the sound waves, but the possibility of individuating a tone happens only after having heard several of the sound impulses and after having judged their frequency. Consequently the sensation of pitch is made possible by previous experiences and cognitions stored in memory, so that one oscillation of the air would not produce a tone.

[2] Cf. [73]. Other considerations on abduction and perception are given in [72].

For Peirce all knowing is *inferring* and inferring is not instantaneous, it happens in a process that needs an activity of comparisons involving many kinds of patterns in a more or less considerable lapse of time. All sensations or perceptions participate in the nature of a unifying hypothesis, that is, in abduction, in the case of emotions too: for example the various sounds made by the musical instruments of the orchestra strike upon the ear, and cause a peculiar musical emotion, completely distinct from the sounds themselves. Emotion is in this case considered by Peirce the same thing as a hypothetic inference, an abduction.

The analogy between abduction and emotion is strong in Peircean writings and Peirce was impressed by the elicitation of an emotion by a complex cognition. Emotion, for example, detects a kind of inconsistency among a set of beliefs, so what beliefs to abandon has to be determined in producing an explanation that resolves the inconsistency. Peirce said emotions are simple predicates that are elicited by complex predicates like in the case of the anxiety that is triggered by the thought that someone has died. In this sense emotions resort to a wonderful example of model-based abductive reasoning, where a signal is sent to the rest of the brain if a certain event occurs. Peirce would have agreed with the current view that in humans, emotions—that can be hardwired or learned through long experience with other human beings and certain situations—are typically non-intentional abductive signals to themselves that "[...] allow humans, who have slender computational resources, to choose among multiple goals, and to act—despite their limited and often incorrect knowledge, and despite their limited physical powers" [44, p. 171].

They can sometimes provide the first indication of an inconsistency, like in the case of anxiety when a loved-one's lateness for an appointment exceeds a certain time but also in high-level cognitive settings such as scientific reasoning.[3] They are also useful to enter mental models of other individuals and tell you about their suitability for possible agreements in the future:

> One woman, for instance, waited for a new colleague in one restaurant, while he sat for over an hour in a different located restaurant in the same chain waiting for her. The fact that he had "stood her up", she said, would be at the back of her mind the next time she had dealings with him. Indeed, it was, even though she stated in her diary that his explanation was convincing. He waited longer than she had in the restaurant, and he was the one who had to phone to find out what had gone wrong. She knew he had been no more in fault than her. Nonetheless the emotion of distrust provided a new kind of forward consistency for her in her relations with this man. This evaluation was compelling even though she held explicit beliefs that were inconsistent with it. The emotion overruled the propositional inconsistency [44, p. 176].

Of course abductive results can also cause emotions, for example depression can be the sign of the invalidity of certain previously abduced hypotheses coupled with some basic aspects of a certain individual's life.

[3] Thagard [69] clearly stresses the central role of the emotion of surprise in finding problems and anomalies in scientific reasoning (and of the emotion of satisfaction caused by a discovery!)

1.2 Iconicity Hybridates Logicality

1.2.1 Knowing as Inferring: The Semiotic View. Is Perception an Inference?

Let us consider some further basic philosophical aspects related to the problem of perception introduced by Peirce. In the following passage, which Peirce decided to skip in his last of the seven Harvard Lectures (14 May 1903), perception is clearly considered a kind of abduction: "A mass of facts is before us. We go through them. We examine them. We find them a confused snarl, an impenetrable jungle. We are unable to hold them in our minds. [...] But suddenly, while we are poring over our digest of the facts and are endeavoring to set them into order, it occurs to us that if we were to assume something to be true that we do not know to be true, these facts would arrange themselves luminously. That is *abduction* [...]".[4] This passage seems to classify abduction as emerging in "perceiving" facts and experiences, and not only in the conclusions of an "inference" [21, pp. 279–280], intended in the classical sense, as expressed by symbols carrying propositional content.

Let us reiterate the following passage, already quoted in the previous subsection; if we say that by perception, knowledge constructions are so instantly reorganized that they become habitual and diffuse and do not need any further testing: "[...] a fully accepted, simple, and interesting inference tends to obliterate all recognition of the uninteresting and complex premises from which it was derived" [49, 7.37]. I also noted: many visual stimuli—that can be considered the "premises" of the involved abduction—are ambiguous, yet people are adept at imposing order on them. Woods comments, suspecting the limitations of the GW-model of abduction[5]: "Perceptual abduction is interesting in a number of ways. As a fast and uncontrolled knowledge production, it operates for the most part automatically and out of sight, so to speak. If true, this puts a good deal of pressure on any suggestion that the GW-schema might be canonical for abduction. In its present formulation, what the schema schematizes is sentential abduction, as Magnani calls it; that is, abduction rendered by symbols carrying propositional content" [76, p. 242]. The semio-philosophical literature on abduction has afforded the conciliation between the sentential and the perceptual aspects of abduction trying to subordinate the second to the first: at a certain level of abstraction, visual stimuli, for example, can be viewed as premisses, and the outputs of perceptual processing—our knowledge that a an obscurely seen face belongs to a friend of ours—in turn be likened to a conjecture derived from the fact or apparent fact that the best causal account of the presence of those stimuli is the presence of our friend. Woods concludes: "In fact, a rather common answer is that what we are told when it is claimed that a certain level of abstraction perception is hypothesis-

[4] Cf. "Pragmatism as the logic of abduction", in [54, pp. 227–241], the quotation is from footnote 12, pp. 531–532.

[5] See below the Appendix: GW and AKM schemas of abduction.

drawing from premisses is that perception is *tacit* hypothesis-drawing from premisses; that the processes that take visual stimuli to a knowledge of birds is abductively inferential in character, but unconsciously and non-symbolically so-*implicitly*".

Certainly the processes that generate perceptual knowledge can be modeled as abduction adopting the subordination of the perceptual side to the sentential one, but I would prefer a coexistence between model-based and sentential aspects, rather than the above conciliation. I endorse a compromise, which I think can increase the intelligibility of abduction as a wide way of inferring hypotheses: given that the concept of abduction is not at all exhausted by the formal models of it, the concept of abduction can be better understood in the light of a composite eco-cognitive view, exactly following the spirit of Peirce's philosophy. If we rigidly separate the two aspects, the inferential one (using the adjective "inferential" just to refer to logical accounts), and the perceptual (as referred to the model-based, or more in general, non sentential semiotic accounts), we rejoin the intellectual conundrum already present in the literature on abduction, caused by the suspected manifest inconsistency of the two views. In this perspective perceptual and inferential views are contrasted and a kind of inconsistency arises, as many researchers contend.

Indeed, it is well-known that in Peirce the inferential side of abduction is initially expressed and denoted by the syllogistic framework. We have just illustrated that, following this point of view the genesis of an—abductive—perceptual judgment would have to be located, following some interpreters, at the level of the premises of the famous Peircean syllogistic schema, that depicts abduction as the fallacy of the affirming the consequent. Moreover, it would be at the level of this perceptual side, and *not* at the level of the logico-sentential one that the proper creative virtues of abduction would be disclosed. The explaining solution would emerge in perceiving facts and experience and not in the conclusion of the logical inference ("the initial conceiving of a novel hypothesis is not the product of an inferential transition" [25, p. 2]).[6]

More simply, as I have anticipated, I think that the two—often considered contrasting—views more simply and coherently can coexist, beyond Peirce,[7] but also in the perspective of the orthodoxy of Peircean texts: the prevailing Peircean *semiotic* conception of inference as a form of sign activity, where the word sign includes "feeling, image, conception, and other representation" offers the solution

[6] It has to be said that some authors (for example [21, p. 280]) contend that, in order to explain abduction as the process of forming an explanatory hypothesis within Peirce's concept of "logic", it is necessary to see both sides as coming together.

[7] It is well-known that in later writings Peirce seems more inclined to see abduction as both insight and inference.

to this potential conflict.[8] In this perspective the meaning of the word inference is not exhausted by its "logical" aspects but is referred to the effect of various sensorial activities. One more reason that supports my contention is that for Peirce the sentential aspects of symbolic disciplines like logic or algebra coexist with model-based features—iconic. Sentential features like symbols and conventional rules are intertwined with the spatial configuration; in Peirce's terms we have already quoted above:

> The truth, however, appears to be that all deductive reasoning, even simple syllogism, involves an element of observation; 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 [49, 3.363].

Perception and Abduction as an Inference to the Best Explanation: The Need of an Eco-Cognitive Model [EC-Model]

In the standard accounts of abductive reasoning, abduction as an inference to the best explanation also involves what Peirce called the inductive/evaluative phase. It is clear that viewing perception as an abduction hardly fits this standard view. No need of empirical evaluation in perception, and consequently it cannot be said that testability is intrinsic to abduction, such as Peirce himself seems to contend in some passages of his writings. In perception the "best abductive choice" is immediately reached—in an uncontrolled way—without the help of an experimental trial (which fundamentally characterizes the received view of abduction in terms of the so-called "inference to the best explanation"). Not only, we have to strongly note that the generation process alone can suffice: in perception the hypothesis generated is immediate and unique.

At the center of my perspective on cognition is the emphasis on the "practical agent", of the individual agent operating "on the ground", that is, in the circumstances of real life. In all its contexts, from the most abstractly logical and mathematical to the most roughly empirical, I always emphasize the cognitive nature of abduction. Reasoning is something performed by cognitive systems. At a certain

[6] It has to be said that some authors (for example [21, p. 280]) contend that, in order to explain abduction as the process of forming an explanatory hypothesis within Peirce's concept of "logic", it is necessary to see both sides as coming together.

[7] It is well-known that in later writings Peirce seems more inclined to see abduction as both insight and inference.

[8] Anderson [4, p. 45] maintains that "Peirce quite explicitly states that abduction is both an insight and an inference. This is a fact to be explained, not to be explained away". Anderson nicely solves this problem by referring to Peirce's theory of the three fundamental categories, Firstness, Secondness, and Thirdness: abduction, as a form of reasoning is essentially a third, but it also occurs at the level of Firstness "as a sensuous form of reasoning" (p. 56 ff.).

level of abstraction and as a first approximation, a cognitive system is a triple (A, T, R), in which A is an *agent*, T is a *cognitive target* of the agent, and R relates to the *cognitive resources* on which the agent can count in the course of trying to meet the target-information, time and computational capacity, to name the three most important. My agents are also *embodied distributed cognitive systems*: cognition is embodied and the interactions between brains, bodies, and external environment are its central aspects. Cognition is occurring taking advantage of a constant exchange of information in a complex distributed system that crosses the boundary between humans, artifacts, and the surrounding environment, where also instinctual and unconscious abilities play an important role. This interplay is especially manifest and clear in various aspects of abductive cognition.[9]

It is in this perspective that we can appropriately consider perceptual abduction—as I have already said—as a fast and uncontrolled knowledge production, that operates for the most part automatically and out of sight, so to speak. This means that—at least in this light—GW-schema is not canonical for abduction, as I have already pointed out. The schema illustrates what I call "sentential abduction" [38, Chap. 1], that is, abduction rendered by symbols carrying propositional content. It is hard to encompass in this model cases of abductive cognition such as perception or the generation of models in scientific discovery.[10] My perspective adopts the wide Peircean philosophical framework, which approaches "inference" *semiotically* (and not simply "*logically*"). It is clear that this semiotic view is considerably compatible with my perspective on cognitive systems as embodied and distributed systems: the GW-schema is instead only devoted to illustrate, even if in a very efficacious way, a subset of the cognitive systems abductive activities, the ones that are performed taking advantage of explicit propositional contents. Woods seems to share this conclusion: "[...] the GW-model helps get us started in thinking about abduction, but it is nowhere close, at any level of abstraction, to running the whole show. It does a good job in modelling the ignorance-preserving character of abduction; but, since it leaves the S_i of the schema's clause (T) unspecified, it makes little contribution to the fill-up problem" [76, p. 244].

In the perspective of my eco-cognitive model (EC-model) the cut-down problem (that is the problem of specifying the conditions for *thinking up* possible candidates for selection) and the fill-up one (that is the problem of finding criteria for hypothesis *selection*) in abductive cognition appear to be spectacularly *contextual*.[11] I lack the space to give this issue appropriate explanation but it suffices

[9] It is interesting to note that recent research on Model Checking in the area of AST (Automated Software Testing) takes advantage of this eco-cognitive perspective, involving the manipulative character of model-based abduction in the practice of adapting, abstracting, and refining models that do not provide successful predictions, cf. [5].

[10] On the knowledge enhancing role of abduction in guessing models in science cf. [41].

[11] Some acknowledgment of the general contextual character of these kinds of criteria, and a good illustration of the role of coherence, unification, explanatory depth, simplicity, and empirical adequacy in the current literature on scientific abductive best explanation, is given in [32].

for the purpose of this study to remember that, for example, one thing is to abduce a model or a concept at the various levels of scientific cognitive activities, where the aim of reaching rational knowledge dominates, another thing is to abduce a hypothesis in literature (a fictional character for example), or in moral reasoning (the adoption/acceptation of a hypothetical judgment as a trigger for moral actions). The case of perception is extreme, because in this case abduction is in itself unconscious and automatic—and immediately "accepted", so to speak—and of course evidentially inert (in the sense that there is no need of the empirical evaluation, instead mandatory in the case of the appropriate activation of a hypothesis in the more composite abductive processes in empirical science). To conclude, the proper experimental test involved in the Peircean evaluation—inductive—phase, which for many researchers would reflect in the most acceptable way the idea of abduction as inference to the best explanation—and so carrier of new reliable knowledge—just constitutes a *special* subclass of the multiple possible modes of adoption/acceptation of an abductive hypothesis.

The backbone of my approach can be found in the manifesto of my eco-cognitive model (EC-model) of abduction in [38]. It might seem awkward to speak of "abduction of a hypothesis in literature," but one of the fascinating aspects of abduction is that not only it can warrant for scientific discovery, but for other kinds of creativity as well. We must not necessarily see abduction as a *problem solving device* that sets off in response to a cognitive irritation/doubt: conversely, it could be supposed that aesthetic abductions (referring to creativity in art, literature, music, etc.) arise in response to some kind of aesthetic irritation that the author (sometimes a *genius*) perceives in herself or in the public. Furthermore, not only aesthetic abductions are free from empirical constraints in order to become the "best" choice: as I am showing throughout this paper, many forms of abductive hypotheses in traditionally-perceived-as-rational domains (such as the setting of initial conditions, or axioms, in physics or mathematics) are relatively free from the need of an empirical assessment. The same could be said of moral judgement: they are eco-cognitive abductions, inferred upon a range of internal and external cues and, as soon as the judgment hypothesis has been abduced, it immediately becomes prescriptive and "true," informing the agent's behavior as such. Assessing that there is a common ground in all of these works of what could be broadly defined as "creativity" does not imply that all of these forms of creativity are the same, contrarily it should spark the need for firm and sensible categorization: otherwise it would be like saying that to construct a doll, a machine-gun and a nuclear reactor are all the same thing because we use our hands in order to do so!

Iconicity and Logicality Intertwined: The Compound Conventional Sign

In another passage, which refers to the "conventional" character of algebraic formulas as icons, the hybridity between sentential and model-based aspects is even clearer and takes advantage of the introduction of the idea of the "compound conventional sign"[12]:

> Particularly deserving of notice are icons in which the likeness is aided by conventional rules. Thus, an algebraic formula is an icon, rendered such by the rules of commutation, association, and distribution of the symbols; that it might as well, or better, be regarded as a compound conventional sign [50, pp. 787 and pp. 26–28 CSP].

It seems for Peirce that iconicity of reasoning, and consequently of abduction are fundamental, like it is clearly stressed in the following further passage: "I said, Abduction, or the suggestion of an explanatory theory, is inference through an Icon" [52, p. 276]. Moreover, induction and deduction are inferences "through an Index" and "through a Symbol" (*ibid.*).

To summarize, it would seem that there is not an inferential aspect of abduction, characterized by the syllogistic model, *separated* from (or contrasted with) the perceptual one, which would be "creative" instead, as many authors contend.[13] I consider the two aspects consistent, and both are perfectly understandable in the framework of Peircean philosophy and semiotics.[14]

A further evidence of the fact that the two aspects of abduction are intertwined derives from the study of children's early word acquisition [61]. Children form knowledge and expectations about the symbolic functioning of a particular word in routine events where model-based perceptual and manipulative aspects of reasoning are predominant and furnish suitable constraints: they generate abductions that help to acquire the content-related symbolic functioning, going beyond what was already experienced. These abduced hypotheses are "practical", about knowing how to use a word to direct attention in a certain way. These hypotheses need not be verbalized by the children, who only later on acquire a more theoretical status through a systematization of their knowledge. It is at this level that they are expressed verbally and concern causal frameworks rather than specific causal mechanisms—for instance of natural kind terms.

To further deepen the particular "inferential" status of abduction we have illustrated above, further problems regarding the relationship between sentential and model-based aspects of abduction have to be analyzed.

[12] Stjernfelt [65] 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.

[13] For example [21, 25].

[14] On the contrary, some authors (for example [21, 22, 46]), as [16, p. 594] synthesized, find a central paradox in "[...] that Peirce holds both that hypotheses are the products of a wonderful imaginative faculty in man and that they are product of a certain sort of logical inference". Furthermore, some commentators seem to maintain that "creative" aspects of abduction would exclusively belong to the perceptual side, as I have already noted above.

1.2.2 Syllogism Versus Perception?

The following is a frequently quoted passage by Peirce on perception and abduction related to the other passage on "perceptual judgment" that I reported above at the beginning of this section:

> Looking out of my window this lovely spring morning I see an azalea in full bloom. No no! I do not see that; though that is the only way I can describe what I see. *That* is a proposition, a sentence, a fact; but what I perceive is not proposition, sentence, fact, but only an image, which I make intelligible in part by means of a statement of fact. This statement is abstract; but what I see is concrete. I perform an abduction when I so much as express in a sentence anything I see. The truth is that the whole fabric of our knowledge is one matted felt of pure hypothesis confirmed and refined by induction. Not the smallest advance can be made in knowledge beyond the stage of vacant staring, without making an abduction in every step.[15]

The classical interpretation of this passage stresses the existence of a vicious circle [21, p. 283]. On the one hand, we learn that the creativity of abduction is based on the genesis of perceptual judgments. On the other hand, it is now said that any perceptual judgment is in itself the result of an abduction. Or, as Peirce says, "[...] our first premises, the perceptual judgments, are to be regarded as an extreme case of abductive inference, from which they differ in being absolutely beyond criticism" [49, 5.181]

Surely it can be maintained that for Peirce perception on the whole is more precisely the act of subsuming sense data or "percepts" under concepts or ideas to give rise to perceptual judgments: we have just said in the previous subsection that he in turn analyzed this act of subsuming as an abductive inference depicted in syllogistic terms

(P1) A well-recognized kind of object, M, has for its ordinary predicates P[1], P[2], P[3], etc.

(P2) The suggesting object, S, has these predicates P[1], P[2], P[3], etc.
(C) Hence, S is of the kind M [49, 8.64].

In this abductive inference—which actually is merely "selective" (see below Sect. 2.3)—the creative act "would" take place in the second premise: if we distinguish in abduction an inferential part and a perceptual one—cf. the previous subsection—(that is the genesis of a perceptual judgment), and if we understand according to Peirce the arising of a perceptual judgment for itself as an abductive inference, then in explaining the possibility of abduction we get an infinite regress. Fortunately, Peirce notes, the "process of forming the perceptual judgment" is "sub-conscious and so not amenable to logical criticism", hence, it is not discrete like sentential inferences, but a "continuous process":

[15] Cf. the article "The proper treatment of hypotheses: a preliminary chapter, toward an examination of Hume's argument against miracles, in its logic and in its history" [1901] (in [50, p. 692]).

> On its side, the perceptive judgment is the result of a process, although of a process not sufficiently conscious to be controlled, or, to state it more truly, not controllable and therefore not fully conscious. If we were to subject this subconscious process to logical analysis, we should find that it terminated in what that analysis would represent as an abductive inference, resting on the result of a similar process which a similar logical analysis would represent to be terminated by a similar abductive inference and so on *ad infinitum*. This analysis would be precisely analogous to which the sophism of Achilles and the Tortoise applied to the chase of the Tortoise by Achilles, and it would fail to represent the real process for the same reason. Namely, just as Achilles does not have to make the series of distinct endeavors which he is represented as making, so this process of forming the perceptual judgment, because it is sub-conscious and so not amenable to logical criticism, does not have to make separate acts of inference, but performs its act in one continuous process [49, 5.181].

This recursiveness, and the related vicious circle, even if stressed by many commentators, do not seem to me really important. I think we can give a simpler explanation of this conflict between the inferential and perceptual side of abduction by recalling once again the Peircean *semiotic* conception of inference as a form of sign activity, where the word sign includes "feeling, image, conception, and other representation".

1.2.3 Explicit, Uncontrolled, and Unconscious Inferences in Multimodal Abduction

As I have maintained in the previous subsections, I think that two contrasting views of abduction such as inferential and model-based (like in the case of perception) can coherently coexist: I have already contended that the prevailing Peircean *semiotic* conception of inference as a form of sign activity offers the solution to the conflict. We also said that for Peirce the sentential aspects of logic, even if central, coexist with model-based features—iconic. Abduction can be performed by words, symbols, and logical inferences, but also by internal processes that treat external sensuous input/signs through merely unconscious mechanisms which give rise to abductive actions and reactions, like in the case of the humble Peircean chicken (cf. below Sect. 2.1) or of human emotions and other various implicit ways of thinking. In these last cases sentential aspects do not play any role (or a dominant role).

We can say, following Thagard [70, 71] that abduction is fundamentally performed in a *multimodal* way: for example, we consciously perform a perceptual judgment about the azalea, and in this case also concepts, ideas and statements certainly play a central abductive role, but—Peirce says, they are only *part* of the whole process: "what I perceived is not proposition, sentence, fact, but only image, which I made intelligible in part by means of a statement of fact".[16] It is in

[16] Cf., again, "The proper treatment of hypotheses: a preliminary chapter, toward an examination of Hume's argument against miracles, in its logic and in its history" (1901) (in [50, p. 692]).

this way that perceptions acquire "meanings": they nevertheless remain "hypotheses" about data we can accept (usually this happens spontaneously) or carefully submit to criticism. It is in this sense that the visual model of perception does not work in isolation from other modes of perception or from other persons or sources of experience [19]. As I have already illustrated in the first subsection above perceptions are withdrawable "inferences", even if not controlled (deliberate), like we control explicit inferences for example in logic and other types of more or less rational human "reasoning" and argumentation.

Being creative is not a peculiarity of perceptual/visual abduction, like—as I have already said—some commentators seem to maintain [25]. Moreover, perception and cognition alike are inherently inferential. If awareness, whether propositional or perceptual, is semiotic, then all awareness involves the interpretation of signs, and all such interpretation is inferential: semiosis not only involves the interpretation of *linguistic* signs, but also the interpretation of *non-linguistic* signs. Abduction of course embraces much of these semiotic performances.

In sum, from a naturalistic perspective both linguistic and non linguistic signs also

1. have an internal semiotic life, as particular configurations of neural networks and chemical distributions (and in terms of their transformations) at the level of human brains, and as somatic expressions,
2. but can also be delegated to many external objects and devices, for example written texts, diagrams, artifacts, etc.

In this "distributed" framework those central forms of abductive cognition that occur in a hybrid way, that is in the interplay between internal and external signs, are of special interest: abduction can be properly seen only in an eco-cognitive framework.

1.2.4 Perception is Semi-Encapsulated

Recent cognitive studies on perception seem to confirm Peirce's philosophical speculations. Through an interdisciplinary approach and suitable experimentation some cognitive scientists (cf. for example Raftopoulos [57, 58]) have recently acknowledged the fact that, in humans, perception (at least in the visual case) is not strictly modular, as Fodor [15] argued, that is, it is not encapsulated, hardwired, domain-specific, and mandatory.[17] Neither is it wholly abductively "penetrable" by higher cognitive states (like desires, beliefs, expectations, etc.), by means of top-down pathways in the brain and by changes in its wiring through perceptual learning, as stressed by Churchland [11]. It is important to consider the three following levels: visual sensation (bodily processes that lead to the formation

[17] Challenges to the modularity hypothesis are illustrated in [42].

of retinal image which are still useless—so to speak—from the high-level cognitive perspective), perception (sensation transformed along the visual neural pathways in a structured representation), and observation, which consists in all subsequent visual processes that fall within model-based/propositional cognition. These processes "[...] include both post-sensory/semantic interface at which the object recognition units intervene as well as purely semantic processes that lead to the identification of the array—high level vision" [58, p. 189].[18]

On the basis of this distinction it seems plausible—as Fodor contends—to think there is a substantial amount of information in perception which is theory-neutral. However, also a certain degree of theory-ladenness is justifiable, which can be seen at work for instance in the case of so-called "perceptual learning". However, this fact does not jeopardize the assumption concerning the basic cognitive impenetrability of perception: in sum, perception is informationally "semi-encaspulated", and also semi-hardwired, but, despite its bottom-down character, it is not insulated—so to speak—from "knowledge". For example, it results from experimentation that illusion is a product of learning from experience, but this does not regard penetrability of perception because these experience-driven changes do not affect a basic core of perception.[19]

Higher cognitive states affect the product of visual modules only after the visual modules "[...] have produced their product, by selecting, acting like filters, which output will be accepted for further processing" [57, p. 434], for instance by selecting through attention, imagery, and semantic processing, which aspects of the retinal input are relevant, activating the appropriate neurons. I have tried to show in this article that I consider these processes essentially abductive, as is also clearly stressed by Shanahan [63], who provides an account of robotic perception from the perspective of a sensory fusion in a unified framework: he describes problems and processes like the incompleteness and uncertainty of basic sensations, top-down information flow and top-down expectation, active perception and attention.[20]

[18] A full treatment of the problem of perception both from a psychological and neural perspective is available in the recent [59]. A recent rich volume that shows the semi-encapsulated character of perception as illustrated by recent cognitive science results is [1].

[19] Evidence on the theory-ladenness of visual perception derived from case-studies in the history of science is illustrated in Brewer and Lambert [10].

[20] Cohn et al. [12] propose a cognitive vision system based on abduction and qualitative spatio-temporal representations capable of interpreting the high level semantics of dynamic scenes. Banerjee [7] presents a computational system able to manage events that are characterized by a large number of individual moving elements, either in pursuit of a goal in groups (as in military operations), or subject to underlying physical forces that group elements with similar motion (as in weather phenomena). Visualizing and reasoning about happenings in such domains are treated through a multilayered abductive inference framework where hypotheses largely flow upwards from raw data to a diagram, but there is also a top-down control that asks lower levels to supply alternatives if the higher level hypotheses are not deemed sufficiently coherent.

It is in this sense that a certain amount of *plasticity* in vision does not imply the full penetrability of perception. As I have already noted, this result does not have to be considered equivalent to the claim that perception is, so to speak, not theory-laden. It has to be acknowledged that even basic perceptual computations obey high-level constraints acting at the brain level, which incorporate implicit and more or less model-based assumptions about the world, coordinated with motor systems. At this level, they lack a semantic content, so as they are not learnt, because they are shared by all, and fundamentally hardwired.

Human auditory perception should also be considered semi-encapsulated [14]. The human auditory system resembles that of other vertebrates, such as mammals, birds, reptiles, amphibians or fish, and it can be thought to derive from simple systems that were originally strictly intertwined with motor systems and thus linked to the sense of space.[21]

Hearing, which works in "dark and cluttered" [14, p. 253] environments, is complementary to other senses, and has both neural bottom-up and top-down characters. The top-down process takes advantage of descending pathways that send *active* information out from a central point and play a part in selectively "listening" to the environment, involving relevant motor aspects (indeed action is fundamental to calibrating perception). The role of hearing in the perception of space is central, complementing multichannel visual information with samples of the acoustic field picked up by the ears: cues to location of source by means of interaural intensity, difference and distance according to cues like loudness are two clear examples of the abductive *inferential* processes performed by hearing that provide substantial models of the scene facing the agent. The whole process is abductive in so far as it provides selections of cues, aggregation of acoustic fragments according to source and an overall hypothetical meaningful explanation of acoustic scenes, that are normally very complex from the point of view of the plurality of acoustic sources. The auditory system of vertebrates which decouples perception from action (motor systems)—still at work together in acoustically rudimentary organisms—enhances economy, speed, and efficacy of the cognitive system by exploiting abstract models of the environment and motor plans.

[21] The example of a simple hypothetical organism equipped with two fins and two eyes [66] can explain this link between perception and action in the case of vision: "The right eye was connected to the left fin by a neuron, and the left eye to the right fin. When a prey appears within the field of the right eye, a command is sent to the left fin to instruct it to move. The organism then turns towards the prey, and this orientation is maintained by bilateral activation until the prey is reached. *Perception* in this primitive organism is not distinct from action" [14, pp. 253–254].

2 Instinct Versus Heuristic Strategies

2.1 Abductive Chickens, Tacit Cognitive Skills, and Manipulative Abduction

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 speak, of spontaneous abduction—analogous to the case of some unconscious/embodied abductive processes in humans:

> 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.[22]

Peirce seems to anticipate the attribution of instinct-based "cognitive" abilities also to animals, of course unconscious and uncontrolled. *Tacit cognitive skills* can be found also in other human cases, even if not related to instinct: what happens when the abductive reasoning in science is strongly related to extra-theoretical actions and manipulations of "external" objects? When abduction is "action-based" on *external models*? When thinking is "through doing", as illustrated in the following simple Peircean example: "A man can distinguish different textures of cloth by feeling: but not immediately, for he requires to move fingers over the cloth, which shows that he is obliged to compare sensations of one instant with those of another" [49, 5.221]. This surely suggests that abductive movements have also interesting extra-theoretical characters and that there is a role in abductive reasoning for various kinds of manipulations of external objects. I would like to reiterate that for Peirce *all* knowing is *inferring* and inferring is not instantaneous, it happens in a process that needs an activity of comparisons involving many kinds of models in a more or less considerable lapse of time. To answer these questions I have delineated the basic features of what I have called *manipulative abduction* [38] by showing how we can find in scientific and everyday reasoning methods of constructivity based on external models and actions, where external things, usually inert from the semiotic point of view, acquire a central cognitive status.

Following the suggestions which come from the studies on embodied and distributed cognition we have to acknowledge the centrality of the so called "disembodiment of the mind", for example in the case of semiotic cognitive processes occurring in science. Disembodiment of the mind refers to the cognitive

[22] Cf. the article "The proper treatment of hypotheses: a preliminary chapter, toward an examination of Hume's argument against miracles, in its logic and in its history" [1901] (in [50, p. 692]).

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. 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 our 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 point out the importance of what I called *manipulative abduction*.

2.1.1 Manipulative Abduction

Manipulative abduction is for example widespread in scientific reasoning, as a process in which a hypothesis is formed resorting to a basically extra-theoretical and extra-sentential behavior that aims at creating communicable accounts of new experiences to the final aim of integrating the successful results 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 for example in a neural engineering laboratory, useful to make observations and "experiments" to transform one cognitive state into another to discover new properties of the target systems, but also when exploiting diagrams and icons. Manipulative abduction refers to those more unplanned and unconscious action-based cognitive processes I have characterized as forms of "thinking through doing".[23] It is clear that manipulative abduction in science basically deals with the handling of external models in their intertwining with the internal ones. Consequently, even if related to experiments occasionally performed with the help of external models sometimes mediated by artifacts, manipulative abduction has to be considered—obviously in mathematics but also in the case of empirical science, evidentially inert, even if of course not necessarily ignorance-preserving, as I have tried to demonstrate in [41].

Finally, let us remember that stressing the role of iconic dimensions of semiosis in the meantime celebrates the virtues of analogy, as a kind of "association by resemblance", as opposed to "association by contiguity":

[23] I have to note that manipulative abduction 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.

> That combination almost instantly flashed out into vividness. Now it cannot be contiguity; for the combination is altogether a new idea. It never occurred to me before; and consequently cannot be subject to any *acquired habit*. It must be, as it appears to be, its analogy, or resemblance in form, to the nodus of my problem which brings it into vividness. Now what can that be but pure fundamental association by resemblance? [49, 7.498].

2.2 *Why Does Abduction Enhance Knowledge? Inference and Instincts*

Peirce provides various justifications of the productive gnoseological role of abduction. They basically resort to the conceptual exploitation of evolutionary and metaphysical ideas, which clearly show that abduction is constitutively akin to truth, certainly ignorance-preserving—because the "absolute truth" is never reached through abduction—but also knowledge enhancing. Peirce himself notes that abductive guesses are belief-inducing and truth making. Not only, it cannot be said that unevidenced belief is itself evidence of malfunction and disorder, and so of falsity.

First of all Peirce considers hypothesis generation a largely instinctual endowment[24] of human beings given by God or related to a kind of Galilean "*lume naturale*": "It is a primary *hypothesis* underlying all abduction that the human mind is akin to the truth in the sense that in a finite number of guesses it will light upon the correct hypothesis" [49, 7.220]. Again, the example of the innate ideas of "every little chicken" is of help to describe this human instinctual endowment:

> How was it that man was ever led to entertain that true theory? You cannot say that it 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 man's 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 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? [49, 5.591].

Paavola [47] illustrates various Peircean ways of understanding the nature of instinct: *naturalistic* (for getting food and for reproducing), *theistic* (also related to the concept of *agapastic* evolution, the idea that the law of love is operative in the cosmos [54, I, pp. 352–371]), and *idealistic*. The last case is related to the so-called

[24] Instinct is of course in part conscious: it is "always partially controlled by the deliberate exercise of imagination and reflection" [49, 7.381].

synechism, according to which everything is continuous and the future will be in some measure continuous with the past, mind and matter are not entirely distinct, and so it is—analogously—for instinct and inference [62, p. 191].[25] The human mind would have been developed under those metaphysical laws that govern the universe so that we can consequently hypothesize that the mind has a tendency to find true hypotheses concerning the universe.

The naturalistic view of instinct involves at least two aspects: *evolutionary/adaptive* and *perceptual*—as a "certain insight" [49, 5.173]: the instinctual insight that leads to a hypothesis is considered by Peirce to be of "the same general class of operations to which Perceptive Judgments belong (*ibid.*) Hence, Peirce considers the capacity to guess correct hypotheses as instinctive and enrooted in our evolution and from this perspective abduction is surely a property of naturally evolving organisms:

> If you carefully consider with an unbiased mind all the circumstances of the early history of science and all the other facts bearing on the question [...] I am quite sure that you must be brought to acknowledge that man's mind has a natural adaptation to imagining correct theories of some kind, and in particular to correct theories about forces, without some glimmer of which he could not form social ties and consequently could not reproduce his kind [49, 5.591].[26]

2.2.1 Synechism: The Blending of Mind and Matter

Peirce also says "Thought is not necessarily connected with brain. It appears in the work of bees, of crystals, and throughout the purely physical world; and one can no more deny that it is really there, than that the colours, the shapes, etc., of objects are really there" [49, 4.551]. It is vital to explain the meaning of this important statement.

[25] I think some of the ideas of the traditional synechism can be usefully deepened in the framework of current research on the so-called multiple realizability thesis, which admits that mind can be "realized" in several material supports, cf. [8, 64].

[26] Cognitive anthropologist Atran advocated a similar view about a century later, arguing in his *Cognitive Foundations of Natural History* that the evolution of religion and pre-scientific forms of knowledge into fully-blown science could be accounted for just recurring to the concepts of *culture* and *cognition*, understanding the latter as "the internal structure of ideas by which the world is conceptualized" [6, p. 3]. Peirce's philosophical speculations have been recently corroborated by a growing interest in *folk science*, that is in the study of uneducated expectations about natural aspects such as biology, mechanics, psychology, physiology and so on. Berlin and his colleagues pioneered the exploration of folkbiological expectations across different cultures [9]. The existence of folk science does not make the case for the actuality of a *lumen naturalis* predisposing humans towards Truth, but for the reality of a penchant (which is also at the level of perception) towards truthfulness: [26] argues that the success of science partially comes from "the ways in which scientists learn to leverage understandings in other minds and to outsource explanatory work through sophisticated methods of deference and simplification of complex systems," (p. 826) but such ways of relying on other people's knowledge in order to achieve better approximations of the truth about a matter are actually preexistent in laypeople and children.

First of all it has to be noted that instincts themselves can undergo modifications through evolution: they are "inherited habits, or in a more accurate language, inherited dispositions" [49, 2.170]. Elsewhere Peirce seems to maintain that instinct is not really relevant in scientific reasoning but that it is typical of just "the reasoning of practical men about every day affairs". So as to say, we can perform instinctive abduction (that is not controlled, not "reasoned") in practical reasoning, but this is not typical of scientific thinking:

> These two [practical and scientific reasoning] would be shown to be governed by somewhat different principles, inasmuch as the practical reasoning is forced to reach some definite conclusion promptly, while science can wait a century or five centuries, if need be, before coming to any conclusion at all. Another cause which acts still more strongly to differentiate the methodeutic of theoretical and practical reasoning is that the latter can be regulated by instinct acting in its natural way, while theory of how one should reason depends upon one's ultimate purpose and is modified with every modification of ethics. Theory is thus at a special disadvantage here; but instinct within its proper domain is generally far keener, and surer, and above all swifter, than any deduction from theory can be. Besides, logical instinct has, at all events, to be employed in applying the theory. On the other hand, the ultimate purpose of pure science, as such, is perfectly definite and simple; the theory of purely scientific reasoning can be worked out with mathematical certainty; and the application of the theory does not require the logical instinct to be strained beyond its natural function. On the other hand, if we attempt to apply natural logical instinct to purely scientific questions of any difficulty, it not only becomes uncertain, but if it is heeded, the voice of instinct itself is that objective considerations should be the decisive ones.[27]

I think that the considerations above do not mean, as some commentators seem to maintain [21, 47, 60], that instinct—as a kind of mysterious, not analyzed, guessing power—"does not" operate at the level of conscious inferences like for example in the case of scientific reasoning. I think a better interpretation is the following that I am proposing here: certainly instinct, which I consider a simple and not a mysterious endowment of human beings, is at the basis of both "practical" and scientific reasoning, in turn instinct shows the obvious origin of both in natural evolution. If every kind of cognitive activity is rooted in a hybrid interplay with external sources and representations, which exhibit their specific constraints and features, it does not appear surprising that "[...] the instincts conducive to assimilation of food, and the instincts conducive to reproduction, must have involved from the beginning certain tendencies to think truly about physics, on the one hand, and about psychics, on the other. It is somehow more than a *mere* figure of speech to say that nature fecundates the mind of man with ideas which, when those ideas grow up, will resemble their father, Nature" [49, 5.591]. Hence, from an evolutionary perspective instincts are rooted in humans in this interplay between internal and external aspects and so it is obvious to see that externalities

[27] Cf. Arisbe Website, http://www.cspeirce.com/menu/library/bycsp/l75/ver1/l75v1-01.htm. The passage comes from MS L75 Logic, regarded as semeiotic (The Carnegie application of 1902).

("Nature") "fecundate" the mind. In this perspective abduction represents the most interesting fruit of this "fecundated" mind.[28]

Beyond the multifarious and sometimes contrasting Peircean intellectual strategies and steps in illustrating concepts like inference, abduction, perception and instinct, which of course are of great interest for the historians of philosophy,[29] the perspective I am describing here seems to me to be able to clearly focus on some central recent cognitive issues which I contend also implicitly underlie Peircean thoughts: nature fecundates the mind because it is through a disembodiment and extension of the mind in nature that in turn nature affects the mind. If we contend a conception of mind as "extended", it is simple to grasp its instinctual part as shaped by evolution through the constraints found in nature itself. It is in this sense that the mind's abductive guesses—both instinctual and reasoned—can be classified as hypotheses "akin to the truth" concerning nature and the external world because the mind grows up together with the representational delegations[30] to the external world that the mind itself has made throughout the history of culture by constructing what some present-day biologists call cognitive niches.[31] In this strict perspective hypotheses are not merely made by pure *unnatural* chance.[32]

Peirce says, in the framework of his *synechism* that "[...] the reaction between mind and matter would be of not essential different kind from the action between parts of mind that are in continuous union" [49, 6.277]. This is clearly seen if we notice that "[...] habit is by no means a mental fact. Empirically, we find that some plants take habits. The stream of water that wears a bed for itself is forming a habit" [49, 5.492]. Finally, here the passage we already quoted above, clearly establishing Peirce's concerns about the mind: "Thought is not necessarily connected with brain. It appears in the work of bees, of crystals, and throughout the purely physical world; and one can no more deny that it is really there, than that the colours, the shapes, etc., of objects are really there" [49, 4.551].

To conclude, seeing abduction as rooted in instinct vs. in inference represents a conflict we can overcome, following Peirce, simply by observing that the work of abduction is partly explicable as an instinctual biological phenomenon and partly as a "logical" operation related to "plastic" cognitive endowments of all

[28] Park [48] compares both Peirce's and my view on instincts and abduction with the estimative power of human and non-human animals, which was one of the internal senses in medieval psychology. In particular he finds amazing analogies with the sophisticated theory of estimative power proposed by Avicenna.

[29] For example, in the latest writings at the beginning of XX century Peirce more clearly stresses the instinctual nature of abduction and at the same time its inferential nature [47, p. 150]. On the various approaches regarding perception in Peircean texts cf. [72].

[30] Representational delegations are those cognitive acts that transform the natural environment in a cognitive one.

[31] Cf. [30, 31, 45]. I have illustrated in detail the concept of cognitive niche in Chap. 6 of [38].

[32] This is not a view that conflicts with the idea of God's creation of human instinct: it is instead meant on this basis, that we can add, with Peirce, the theistic hypothesis, if desired.

organisms. I entirely agree with Peirce: in a naturalistic perspective, a guess in science, the appearance of a new hypothesis, is also[33] a biological phenomenon and so it is related to instinct: in the sense that first of all we can analogize the appearance of a new hypothesis to a "trustworthy" chance variation in biological evolution [49, 7.38], even if of course the evolution—for example—of scientific guesses does not conform to the pattern of biological evolution [13, p. 427]. An abduced hypothesis introduces a change (and a chance) in the semiotic processes to advance new perspectives in the co-evolution of the organism and the environment: it is in this way that they find a continuous mutual variation. The organism modifies its character in order to reach better fitness; however, the environment (already artificially—culturally—modified, i.e. a cognitive niche), is equally continuously changing and very sensitive to every modification. In summary, the fact that abduction is akin to truth is guaranteed at both the metaphysical and evolutionary levels: the case of instinct and the case of perception described by Peirce are striking, both provide abductions that are immediately and spontaneously generated but at the same time activated and efficacious, certainly not "in sufferance" (as Woods would say, referring to the case of the standard activity of abducing hypotheses in natural science), and so in need of empirical evaluation.

2.3 Peircean Chickens, Human Agents, Logical Agents

It is certainly true that Peirce is also convinced that there is a gap between logic and scientific reasoning on one side and practical reasoning on the other (he also rejects the possibility of a practical "logic" and consequently of a logic of abductive reasoning in practical contexts): "In everyday business reasoning is tolerably successful but I am inclined to think that it is done as well without the aid of theory as with it" [55, p. 109]. "My proposition is that logic, in the strict sense of the term, has nothing to do with how you think [...]. Logic in the narrower sense is that science which concerns itself primarily with distinguishing reasonings into good and bad reasonings, and with distinguishing probable reasonings into strong and weak reasonings. Secondarily, logic concerns itself with all that it must study in order to draw those distinctions about reasoning, and with nothing else" (*ibid.*, p. 143). We have illustrated that the role of instinct at the level of human unconscious reasoning is obvious, this kind of cognition has been wired by the evolution (like it also happens in the case of some animals, for example the Peircean chicken above), and in some organisms cannot be even partially accessed by consciousness.

[33] Of course this conclusion does not mean that artifacts like computers do not or cannot perform abductions. The recent history of artificial intelligence in building systems able to perform diagnoses and creativity clearly illustrates this point.

However, today we have at our disposal many logics of abduction: Gabbay and Woods [17] contend that these logics are just formal and somewhat idealized descriptions of an abductive agent. A real human agent (the every day "business reasoner") can be considered a kind of biological realization of a nonmonotonic paraconsistent base logic and surely the strategies provided by classical logic and some strictly related non standard logics form a very small part of individual cognitive skills, given the fact that human agents are not in general dedicated to error avoidance like "classical" logical agents [37]. The fact that human beings are error prone does not have to be considered as something bad from the evolutionary point of view, as also Quine contends [56]: for example, hasty generalizations (and many other fallacies) are bad ways of reasoning but can be the best means for survival (or at least for reaching good aims) in particular contexts [33, 74].

Questions of relevance and plausibility regarding the activity of guessing hypotheses (in an inductive or abductive way) are embedded at the level of the implicit reasoning performances of the practical agent. It will be at the level of the formal model, as an idealized description of what an abductive/inductive agent does, that for example questions of economy, relevance and plausibility, in so far as they can be rendered in terms of heuristics strategies, will be explicitly described.[34]

In summary, from a semiotic point of view, the idea that there is a conflict (or a potential conflict) (present in Peirce's texts too) between views of abduction—and of practical reasoning—in terms of heuristic strategies or in terms of instinct (insight, perception) [21, 46, 47], appears to me old-fashioned. The two aspects simply coexist at the level of the real organic agent, it depends on the cognitive/semiotic perspective we adopt:

1. we can see it as a practical agent that mainly takes advantage of its implicit endowments in terms of guessing right, wired by evolution, where of course instinct or hardwired programs are central, or
2. we can see it as the user of explicit and more or less abstract semiotic devices (language, logic, visualizations, etc.) internally stored and/or externally available—hybrid—where heuristic plastic strategies are at work. These strategies, used for guessing right, exploit various and contextual relevance and plausibility criteria built up during cultural evolution and made available in cognitive niches, where they can also potentially be taught and learnt.

What is still important to note is that these heuristic strategies and reasoning devices are determined and created at the level of individuals and through the interplay of both the internal and the external agency already endowed with those cognitive delegated representations and tools occurring in the continuous semiotic activity of "disembodiment of mind" that I have illustrated in Chap. 3 of [38].

[34] On the role of strategies, plausibility, and economy of research and their relationships with Peircean Grammar, Critic, and Methodeutic cf. [46]. A detailed and in-depth description of these difficult aspects of philosophical and semiotic issues of Peirce's approach is given in [28].

This interplay of course occurs both at the contextual level of learning in the individual history and at the level of the evolutionary effects. The efficiency of these strategies in terms of "naturalistic"—and so instinctual—characters is just guaranteed by this interplay.

In the first case the role of instinct is clear in the sense that the cognitive skills have been wired thanks to the evolutionary interplay between organic agents and their environments. In the second case the role of instinct is still at stake, but in so far as it is at the origins of the historical process of formation of heuristic strategies, and thus of reasoning devices, there is the same interplay between organism and environment, internal and external representations, but in terms of explicit tools sedimented in historical practices and learnt—and possibly improved—by the individuals. Sedimented heuristics are *context-dependent*, to make an example, scientists, in their various fields, make use of many heuristic strategies that are explicitly stated, learnt, and enhanced. These reasoning processes, even if far from being considered merely instinctual, still serve what Peirce calls "the probable perpetuation" [54, II, pp. 464 f.] of the race.

For example, when we model abduction through a computational logic-based system, the fundamental operation is to search, which expresses the heuristic strategies [68]. When there is a problem to solve, we usually face several possibilities (hypotheses) and we have to select the suitable one (selective abduction). Accomplishing the assigned task requires that we have to search through the whole space of potential solutions to find the desired one. In this situation we have to rely on heuristics, that are rules of thumb expressed in sentential terms. The well-known concept of *heuristic search*, which is at the basis of many computational systems based on propositional rules, can perform this kind of sentential abduction (selective). We have to point out that other computational tools can be used to this aim, like neural and probabilistic networks, and frames-like representations also able to imitate both sentential, model-based, and hybrid ways of reasoning of real human agents, but less appropriate to model the traditional concept of heuristic strategy.

3 Conclusion

I this article I have illustrated that, to understand abduction, an "archeological"—and at the same time interdisciplinary—effort is mandatory, which takes advantage of both the critical revision of philosophical classical speculations and recent epistemological and cognitive results. To this aim I have analyzed some "canonic" aspects of Peirce's philosophy resorting to the description of the role of abduction in inferences, in perception, in diagrams and icons, and as instinct-based. I have further intertwined these traditional issues to the the recent analysis of creative, selective, model-based, multi-modal, and manipulative abduction, adopting an extended and rich eco-cognitive perspective (EC-model). Following this intellectual route, *understanding abduction* also becomes a way of better recognizing the limitations of formal and computational models, otherwise so useful to focus on other relevant aspects of abductive reasoning, such as ignorance-preservation, relevance and plausibility criteria, and the problem of the inference to the best explanation. Peircean analysis helps us to better grasp how sentential, model-based, and manipulative aspects of abduction have to be seen as intertwined, and indispensable for building a satisfactory and unified model of abduction.

Acknowledgments For the instructive criticisms and precedent discussions and correspondence that helped me to develop my critique of the ignorance preserving character of abduction, and my analysis of the related model-based aspects, I am indebted and grateful to John Woods, Mauricio Suárez, Shahid Rahman, Alirio Rosales, and Tommaso Bertolotti. In writing the current article, I have excerpted, revised, and integrated some paragraphs of my book *Abductive Cognition*.

Appendix: GW and AKM Schemas of Abduction

I have already said that the GW-model[35] does a good job in modeling the ignorance-preserving character of abduction and—I am convinced—in designing the correct intellectual framework we should adopt especially when dealing with the problem of abduction as an inference to the best hypothesis/explanation. Following Gabbay and Woods' contention, it is clear that "[...] abduction is a procedure in which something that lacks epistemic virtue is accepted because it has virtue of another kind" [17, p. 62].

For example: "Let *S* be the standard that you are not able to meet (e.g., that of mathematical proof). It is possible that there is a lesser epistemic standard S' (e.g., having reason to believe) that you do meet" [77, Chap. 10]. Focusing attention on this cognitive aspect of abduction, and adopting a logical framework centered on practical agents, Gabbay and Woods [17] contend that abduction (basically seen as a *scant-resource* strategy, which proceeds in absence of knowledge) presents an *ignorance-preserving* (or, better, an *ignorance mitigating*) character. Of course "[...] it is not at all necessary, or frequent, that the abducer be wholly in the dark, that his ignorance be total. It needs not be the case, and typically isn't, that the abducer's choice of a hypothesis is a blind guess, or that nothing positive can be said of it beyond the role it plays in the subjunctive attainment of the abducer's original target (although sometimes this is precisely so)" (cit.). In this perspective, abductive reasoning is a *response* to an ignorance-problem: one has an ignorance-problem when one has a cognitive target that cannot be attained on the basis of what one currently knows. Ignorance problems trigger one or other of three responses. In the first case, one overcomes one's ignorance by attaining some additional knowledge (subduance). In the second instance, one yields to one's ignorance (at least for the time being) (surrender). In the third instance, one abduces [77, Chap. 10] and so has some positive basis for new action even if in the presence of the constitutive ignorance.

[35] That is Gabbay and Woods Schema.

From this perspective the general form of an abductive inference can be symbolically rendered as follows. Let α be a proposition with respect to which you have an ignorance problem. Putting T for the agent's epistemic target with respect to the proposition α at any given time, K for his knowledge-base at that time, K^* for an immediate accessible successor-base of K that lies within the agent's means to produce in a timely way,[36] R as the attainment relation for T, $\rightsquigarrow$ as the *subjunctive* conditional relation, H as the agent's hypothesis, $K(H)$ as the revision of K upon the addition of H, $C(H)$ denotes the conjecture of H and H^c its activation. The general structure of abduction can be illustrated as follows (GW-schema):

1. $T!\alpha$	[setting of T as an epistemic target with respect to a proposition α]
2. $\neg(R(K,T)$	[fact]
3. $\neg(R(K^*,T)$	[fact]
4. $H \notin K$	[fact]
5. $H \notin K^*$	[fact]
6. $\neg R(H,T)$	[fact]
7. $\neg R(K(H),T)$	[fact]
8. If $H \rightsquigarrow R(K(H),T)$	[fact]
9. H meets further conditions $S_1, \ldots. S_n$	[fact]
10. Therefore, $C(H)$	[sub-conclusion, 1–9]
11. Therefore, H^c	[conclusion, 1–10]

It is easy to see that the distinctive epistemic feature of abduction is captured by the schema. It is a given that H is not in the agent's knowledge-set. Nor is it in its immediate successor. Since H is not in K, then the revision of K by H is not a knowledge-successor set to K. Even so, $H \rightsquigarrow (K(H),T)$. So we have an ignorance-preservation, as required (cf. [77, Chap. 10]).

[*Note*: Basically, line 9. indicates that H has no more plausible or relevant rival constituting a greater degree of subjunctive attainment. Characterizing the S_i is the most difficult problem for abductive cognition, given the fact that in general there are many possible candidate hypotheses. It involves for instance the *consistency* and *minimality* constraints.[37] These constraints correspond to the

[36] K^* is an accessible successor of K to the degree that an agent has the know-how to construct it in a timely way; i.e., in ways that are of service in the attainment of targets linked to K. For example if I want to know how to spell 'accommodate', and have forgotten, then my target can't be hit on the basis of K, what I now know. But I might go to my study and consult the dictionary. This is K^*. It solves a problem originally linked to K.

[37] I have shown in this article that, in the case of inner processes in organic agents, this subprocess—here explicitly modeled thanks to a formal schema—is considerably implicit, and so also linked to unconscious ways of inferring, or even, in Peircean terms, to the activity of the instinct [49, 8.223] and of what Galileo called the *lume naturale* [49, 6.477], that is the innate fair for guessing right. This and other cognitive aspects can be better illustrated thanks to the alternative EC-model model of abduction I have introduced in this article.

lines 4 and 5 of the standard AKM schema of abduction,[38] which is illustrated as follows:

where of course the conclusion operator $\looparrowright$ cannot be classically interpreted].[39]

1. E
2. $K \not\looparrowright E$
3. $H \not\looparrowright E$
4. $K(H)$ is consistent
5. $K(H)$ is minimal
6. $K(H) \looparrowright E$
7. Therefore, H.

[17, pp. 48–49].

Finally, in the GW-schema $C(H)$ is read "It is justified (or reasonable) to conjecture that H" and H^c is its activation, as the basis for *planned* "actions".

In sum, in the GW-schema T cannot be attained on the basis of K. Neither can it be attained on the basis of any successor K^* of K that the agent knows then and there how to construct. H is not in K: H is a hypothesis that when reconciled to K produces an updated $K(H)$. H is such that if it were true, then $K(H)$ would attain T. The problem is that H is *only hypothesized*, so that the truth is not assured. Accordingly Gabbay and Woods contend that $K(H)$ *presumptively* attains T. That is, having hypothesized that H, the agent just "presumes" that his target is now attained. Given the fact that presumptive attainment is not attainment, the agent's abduction must be considered as preserving the ignorance that already gave rise to her (or its, in the case for example of a machine) initial ignorance-problem. Accordingly, abduction does not have to be considered the "solution" of an ignorance problem, but rather a response to it, in which the agent reaches presumptive attainment rather than actual attainment. $C(H)$ expresses the conclusion that it follows from the facts of the schema that H is a worthy object of conjecture. It is important to note that in order to solve a problem it is not necessary that an agent actually conjectures a hypothesis, but it is necessary that she states that the hypothesis is *worthy of conjecture*.

[38] The classical schematic representation of abduction is expressed by what [17] call AKM-schema, which is contrasted to their own (GW-schema), which I am just explaining in this subsection. For A they refer to Aliseda [2, 3], for K to Kowalski [27], Kuipers [29], and Kakas *et al.* [23], for M to Magnani [36] and Meheus [43]. A detailed illustration of the AKM schema is given in [Magnani (2009), Chap. 2, Sect. 2.1.3].

[39] The target has to be an explanation and $K(H)$ bears R^{pres} [that is the relation of presumptive attainment] to T only if there is a proposition V and a consequence relation $\looparrowright$ such that $K(H) \looparrowright V$, where V represents a *payoff proposition* for T. In turn, in this schema explanations are interpreted in consequentialist terms. If E is an explanans and E' an explanandum the first explains the second only if (some authors further contend if and only if) the first implies the second. It is obvious to add that the AKM schema embeds a D-N (deductive-nomological) interpretation of explanation, as I have already stressed in [36, p. 39].

Finally, considering H justified to conjecture is not equivalent to considering it justified to accept/activate it and eventually to send H to experimental trial. H^c denotes the *decision* to release H for further premissory work in the domain of enquiry in which the original ignorance-problem arose, that is the activation of H as a positive *cognitive* basis for action. Woods usefully observes:

> There are lots of cases in which abduction stops at line 10, that is, with the conjecture of the hypothesis in question but not its activation. When this happens, the reasoning that generates the conjecture does not constitute a positive basis for new action, that is, for acting *on* that hypothesis. Call these abductions *partial* as opposed to full. Peirce has drawn our attention to an important subclass of partial abductions. These are cases in which the conjecture of H is followed by a decision to submit it to experimental test. Now, to be sure, doing this is an action. It is an action *involving* H but it is not a case of acting *on* it. In a full abduction, H is activated by being released for inferential work in the domain of enquiry within which the ignorance-problem arose in the first place. In the Peircean cases, what counts is that H is withheld from such work. Of course, if H goes on to test favourably, it may then be released for subsequent inferential engagement [75].

We have to remember that this process of evaluation and so of activation of the hypothesis, is not abductive, but inductive, as Peirce contended. Woods adds: "Now it is quite true that epistemologists of a certain risk-averse bent might be drawn to the admonition that partial abduction is as good as abduction ever gets and that complete abduction, inference-activation and all, is a mistake that leaves any action prompted by it without an adequate rational grounding. This is not an unserious objection, but I have no time to give it its due here. Suffice it to say that there are real-life contexts of reasoning in which such conservatism is given short shrift, in fact is ignored altogether. One of these contexts is the criminal trial at common law" [75].

In the framework of the GW-schema it cannot be said that testability is intrinsic to abduction, such as it is instead maintained in the case of some passages of Peirce's writings.[40] This activity of testing, I repeat, which in turn involves degrees of risk proportioned to the strength of the conjecture, is strictly cognitive/epistemic and inductive in itself, for example an experimental test, and it is an intermediate step to release the abduced hypothesis for inferential work in the domain of enquiry within which the ignorance-problem arose in the first place.

Through abduction the basic ignorance—that does not have to be considered total "ignorance"—is neither solved nor left intact: it is an ignorance-preserving accommodation of the problem at hand, which "mitigates" the initial cognitive "irritation" (Peirce says "the irritation of doubt").[41] As I have already stressed, in a defeasible way, further action can be triggered either to find further abductions or

[40] When abduction stops at line 10., the agent is not prepared to accept $K(H)$, because of supposed adverse consequences.

[41] "The action of thought is excited by the irritation of doubt, and ceases when belief is attained; so that the production of belief is the sole function of thought" [53, p. 261].

to “solve” the ignorance problem, possibly leading to what the “received view” has called the *inference to the best explanation* (IBE).

It is clear that in the framework of the GW-schema the inference to the best explanation—if considered as a truth conferring achievement justified by the empirical approval—cannot be a case of abduction, because abductive inference is constitutively ignorance-preserving. In this perspective the inference to the best explanation involves the generalizing and evaluating role of *induction*. Of course it can be said that the requests of originary thinking are related to the depth of the abducer’s ignorance.

In [41] I have extensively analyzed and criticized the ignorance-preserving character of abduction, taking advantage of my *eco-cognitive model* (EC-model) of abduction and of three examples taken from the areas of both philosophy and epistemology. Indeed, through abduction, knowledge can be enhanced, even when abduction is not considered an inference to the best explanation in the classical sense of the expression, that is an inference necessarily characterized by an empirical evaluation phase, or an inductive phase, as Peirce called it. Hence, abduction is not always ignorance-preserving, but knowledge enhancing.

Finally, let us reiterate a passage taken from Woods’ quotation above: “There are lots of cases in which abduction stops at line 10, that is, with the conjecture of the hypothesis in question but not its activation. When this happens, the reasoning that generates the conjecture does not constitute a positive basis for new action, that is, for acting *on* that hypothesis”. We do not have to forget that, as I have illustrated in this paper, various ways of *positively* enhancing knowledge are occurring also in the case of evidentially inert abductions (perception, instinct, scientific models [41], etc.), and very often a human abductive guess is activated and becomes a basis for action even if it has provided absolutely unreliable—if seen in the light of positive rational criteria of acceptation—knowledge. This is the case for example of the role of abductive guesses in the so-called fallacious and other kinds of reasoning, where the simple struggle that is occurring at the level of the so-called *coalition enforcement* is at stake.[42]

References

1. Albertazzi, L., van Tonder, G.J., Vishwanath, D. (eds.): Perception Beyond Inference: The Information Content of Visual Processes. The MIT Press, Cambridge (2011)
2. Aliseda, A.: Seeking explanations: abduction in logic, philosophy of science and artificial intelligence. PhD thesis, Institute for Logic, Language and Computation, Amsterdam (1997)
3. Aliseda, A.: Abductive Reasoning: Logical Investigations into Discovery and Explanation. Springer, Berlin (2006)
4. Anderson, D.R.: Creativity and the Philosophy of Charles S. Peirce. Claredon Press, Oxford (1987)

[42] I have analyzed the role of abduction in coalition enforcement, as a cognitive tool of the so-called *military intelligence* in [39] and, in the case of *epistemic warfare*, in [40].

5. Angius, N.: Towards model-based abductive reasoning in automated software testing. Logic J. IGPL (2013, Forthcoming), doi:10.1093/jigpal/jzt006
6. Atran, S.: Cognitive Foundations of Natural History: Towards an Anthropology of Science. Cambridge University Press, Cambridge (1990)
7. Banerjee, B.: A layered abductive inference framework for diagramming group motions. Logic J. IGPL **14**(2), 363–378 (2006)
8. Baum, E.B.: What is Thought?. The MIT Press, Cambridge (2006)
9. Berlin, B., Breedlove, D., Raven, P.: General principles of classification and nomenclature in folk biology. Am. Anthropol. **74**, 214–242 (1973)
10. Brewer, W.F., Lambert, B.L.: The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. Philos. Sci. **68**, S176–S186 (2001). Proceedings of the PSA 2000 Biennal Meeting
11. Churchland, P.M.: Perceptual plasticity and theoretical neutrality: a reply to Jerry Fodor. Philos. Sci. **55**, 167–187 (1988)
12. Cohn, A.G., Magee, D.R., Galata, G., Hogg, D.C., Hazarika, S.M.: Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. In: Freska, C., Habel, C., Wender, K.F. (eds.) Spatial Cognition III, pp. 232–248. Springer, Berlin (2002)
13. Colapietro, V.: Conjectures concerning an uncertain faculty. Semiotica **153**(1/4), 413–430 (2005)
14. de Cheveigné, A.: Hearing, action, and space. In: Andler, D., Ogawa, Y., Okada, M., Watanabe, S. (eds.) Reasoning and Cognition, pp. 253–264. Keio University Press, Tokyo (2006)
15. Fodor, J.: Observation reconsidered. Philos. Sci. 51:23–43 (1984). Reprinted in [18, pp. 119–139]
16. Frankfurt, H.: Peirce's notion of abduction. J. Philos. **55**, 593–597 (1958)
17. Gabbay, D.M., Woods, J.: The Reach of Abduction. A Practical Logic of Cognitive Systems, vol. 2. North-Holland, Amsterdam (2005)
18. Goldman, A.I. (ed.): Readings in Philosophy and Cognitive Science. Cambridge University Press, Cambridge (1993)
19. Gooding, D.: Creative rationality: towards an abductive model of scientific change. Philosophica **58**(2), 73–102 (1996)
20. Hintikka, J.: What is abduction? The fundamental problem of contemporary epistemology. Trans. Charles S. Peirce Soc. **34**, 503–533 (1998)
21. Hoffmann, M.H.G.: Problems with Peirce's concept of abduction. Found. Sci. **4**(3), 271–305 (1999)
22. Hoffmann, M.H.G.: How to get it. Diagrammatic reasoning as a tool for knowledge development and its pragmatic dimension. Found. Sci. **9**, 285–305 (2004)
23. Kakas, A., Kowalski, R.A., Toni, F.: Abductive logic programming. J. Logic Comput. **2**(6), 719–770 (1993)
24. Kant, I.: Critique of Pure Reason (Translated by N. Kemp Smith, originally published 1787, reprint 1998). MacMillan, London (1929)
25. Kapitan, T.: Peirce and the structure of abductive inference. In: Houser, N., Roberts, D.D., van Evra, J. (eds.) Studies in the Logic of Charles Sanders Peirce, pp. 477–496. Indiana University Press, Bloomington and Indianapolis (1997)
26. Keil, F.: The feasibility of folk science. Cogn. Sci. **34**, 826–862 (2010)
27. Kowalski, R.A.: Logic for Problem Solving. Elsevier, New York (1979)
28. Kruijff, G.-J.-M.: Peirce's late theory of abduction: a comprehensive account. Semiotica **153**(1/4), 431–454 (2005)
29. Kuipers, T.A.F.: Abduction aiming at empirical progress of even truth approximation leading to a challenge for computational modelling. Found. Sci. **4**, 307–323 (1999)
30. Laland, K.N., Odling-Smee, F.J., Feldman, M.W.: Niche construction, biological evolution and cultural change. Behav. Brain Sci. **23**(1), 131–175 (2000)

31. Laland, K.N., Odling-Smee, F.J., Feldman, M.W.: Cultural niche construction and human evolution. J. Evol. Biol. **14**, 22–33 (2001)
32. Mackonis, A.: Inference to the best explanation, coherence and other explanatory virtues. Synthese **190**, 975–995 (2013)
33. Magnani, L., Belli, E.: Agent-based abduction: being rational through fallacies. In: Magnani, L. (ed.) Model-Based Reasoning in Science and Engineering. Cognitive Science, Epistemology, Logic, pp. 415–439. College Publications, London (2006)
34. Magnani, L., Civita, S., Previde Massara, G.: Visual cognition and cognitive modeling. In: Cantoni, V. (ed.) Human and Machine Vision: Analogies and Divergences, pp. 229–243. Plenum Publishers, New York (1994)
35. Magnani, L.: Visual abduction: philosophical problems and perspectives. In: AAAI Spring Symposium, pp. 21–24. American Association for Artificial Intelligence, Stanford, CA: Comment to R. Lindsay, Generalizing from diagrams (1996)
36. Magnani, L.: Abduction, Reason, and Science: Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
37. Magnani, L.: Abduction and cognition in human and logical agents. In: Artemov, S., Barringer, H., Garcez, A., Lamb, L., Woods, J. (eds.) We Will Show Them: Essays in Honour of Dov Gabbay, vol. II, pp. 225–258. College Publications, London (2007)
38. Magnani, L.: Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Heidelberg (2009)
39. Magnani, L.: Understanding Violence. The Interwining of Morality, Religion, and Violence: A Philosophical Stance. Springer, Heidelberg (2011)
40. Magnani, L.: Scientific models are not fictions. Model-based science as epistemic warfare. In: Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science, Western and Eastern Studies, pp. 1–38. Springer, Heidelberg (2012)
41. Magnani, L.: Is abduction ignorance-preserving? Conventions, models, and fictions in science. Logic J. IGPL (2013, Forthcoming), doi:10.1093/jigpal/jzt012
42. Marcus, G.F.: Cognitive architecture and descent with modification. Cognition **101**, 443–465 (2006)
43. Meheus, J., Verhoeven, L., Van Dyck, M., Provijn, D.: Ampliative adaptive logics and the foundation of logic-based approaches to abduction. In: Magnani, L., Nersessian, N.J., Pizzi, C. (eds.) Logical and Computational Aspects of Model-Based Reasoning, pp. 39–71. Kluwer Academic Publishers, Dordrecht (2002)
44. Oatley, K., Johnson-Laird, P.N.: Emotion and reasoning to consistency. In: Moore, S.C., Oaksford, M. (eds.) Emotional Cognition, pp. 157–181. Johns Benjamins, Amsterdam/Philadelphia (2002)
45. Odling-Smee, F.J., Laland, K.N., Feldman, M.W.: Niche Construction. The Neglected Process in Evolution. Princeton University Press, Princeton (2003)
46. Paavola, S.: Abduction through grammar, critic and methodeutic. Trans. Charles S. Peirce Soc. **40**(2), 245–270 (2004)
47. Paavola, S.: Peircean abduction: instinct or inference? Semiotica **153**(1/4), 131–154 (2005)
48. Park, W.: Abduction and estimation in animals. Found. Sci. **17**(4), 321–337 (2012)
49. Peirce, C.S.: 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).
50. Peirce, C.S.: The Charles S. Peirce Papers: Manuscript Collection in the Houghton Library. The University of Massachusetts Press, Worcester (1966). Annotated Catalogue of the Papers of Charles S. Peirce. Numbered according to Richard S. Robin. Available in the Peirce Microfilm edition. Pagination: CSP = Peirce / ISP = Institute for Studies in Pragmaticism.
51. Peirce, C.S.: The New Elements of Mathematics by Charles Sanders Peirce. Mouton/Humanities Press, The Hague-Paris/Atlantic Higlands (1976). (vols I-IV, edited by C. Eisele)
52. Peirce, C.S.: Pragmatism as a Principle and Method of Right Thinking. The 1903 Harvard Lectures on Pragmatism. State University of New York Press, Albany (1986). (Turrisi, P.A., Peirce, C.S. (ed.) Lectures on Pragmatism. Cambridge, March 26 - May 17 (1903). Reprinted in [54, II, pp. 133–241])

53. Peirce, C.S.: Historical Perspectives on Peirce's Logic of Science: a History of Science. Mouton, Berlin (1987). (vols. I-II, edited by C. Eisele)
54. Peirce, C.S.: The Essential Peirce. Selected Philosophical Writings. Indiana University Press, Bloomington and Indianapolis (1992–1998). (Vol. 1 (1867–1893), ed. by N. Houser & C. Kloesel; vol. 2 (1893–1913) ed. by the Peirce Edition Project)
55. Peirce, C.S.: Reasoning and the Logic of Things: the 1898 Cambridge Conferences Lectures by Charles Sanders Peirce. Harvard University Press, Amsterdam (2005) (Edited by K. L. Ketner)
56. Quine, W.V.O.: Natural kinds. In Ontological Relativity and Other Essays. Columbia University Press, New York (1969)
57. Raftopoulos, A.: Is perception informationally encapsulated? The issue of theory-ladenness of perception. Cogn. Sci. **25**, 423–451 (2001)
58. Raftopoulos, A:. Reentrant pathways and the theory-ladenness of perception. Philos. Sci. **68**, S187–S189 (2001). Proceedings of PSA 2000 Biennal Meeting
59. Raftopoulos, A:. Cognition and Perception. How Do Psychology and Neural Science Inform Philosophy? The MIT Press, Cambridge (2009)
60. Rescher, N.: Peirce on abduction, plausibility, and efficiency of scientific inquiry. In: Rescher, N. (ed.) Essays in the History of Philosophy, pp. 309–326. Avebury, Aldershot (1995)
61. Roberts, L.D.: The relation of children's early word acquisition to abduction. Found. Sci. **9**(3), 307–320 (2004)
62. Santaella, L.: Abduction: the logic of guessing. Semiotica **153**(1/4), 175–198 (2005)
63. Shanahan, M.: Perception as abduction: turning sensory data into meaningful representation. Cogn. Sci. **29**, 103–134 (2005)
64. Shapiro, L.A.: The Mind Incarnate. The MIT Press, Cambridge (2004)
65. Stjernfelt, F.: Diagrammatology. An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics. Springer, Berlin/New York (2007)
66. Szentagothai, J., Arbib, M.A.: Conceptual Models of Neural Organization. The MIT Press, Cambridge (1975)
67. Thagard, P.: Computational Philosophy of Science. The MIT Press, Cambridge (1988)
68. Thagard, P.: Mind. Introduction to Cognitive Science. The MIT Press, Cambridge (1996)
69. 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)
70. Thagard, P.: How does the brain form hypotheses? Towards a neurologically realistic computational model of explanation. In: Thagard, P., Langley, P., Magnani, L., Shunn, C. (eds.) Symposium Generating explanatory hypotheses: mind, computer, brain, and world. Cognitive Science Society, CD-Rom. Proceedings of the 27th International Cognitive Science Conference. Stresa, Italy (2005).
71. Thagard, P.: Abductive inference: from philosophical analysis to neural mechanisms. In: Feeney, A., Heit, E. (eds.) Inductive Reasoning: Experimental, Developmental, and Computational Approaches, pp. 226–247. Cambridge University Press, Cambridge (2007)
72. Tiercelin, C.: Abduction and the semiotic of perception. Semiotica **153**(1/4), 389–412 (2005)
73. Turrisi, P.A.: Peirce's logic of discovery: abduction and the universal categories. Trans. Charles S. Peirce Soc. **26**, 465–497 (1990)
74. Woods, J.: The Death of Argument. Kluwer Academic Publishers, Dordrecht (2004)
75. Woods, J.: Ignorance, inference and proof: abductive logic meets the criminal law. In: Tuzet, G., Canale, D. (eds.) The Rules of Inference: Inferentialism in Law and Philosophy, pp. 151–185. Egea, Heidelberg (2009)
76. Woods, J.: Recent developments in abductive logic. Stud. Hist. Philos. Sci. **42**(1):240–244 (2011). Essay Review of Magnani, L.: Abductive Cognition. The Epistemologic and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Heidelberg (2009)
77. Woods, J.: Errors of Reasoning. Naturalizing the Logic of Inference. College Publications, London (2013, Forthcoming)

How to Learn Abduction from Animals? From Avicenna to Magnani

Woosuk Park

Abstract Magnani's recent discussion of animal abduction sheds considerable light on both instinctual and inferential character of Peircean abduction. Inspired by this, I elsewhere noted some analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. Also, I confirmed the dividing role and function of the Beast-Machine controversy in the history of the study of animal cognition. In this paper, I propose to discuss rather extensively some of the most salient differences between Avicenna and Peirce-Magnani. Unlike estimation that only allows animals to sense what is insensible, i.e., intentions, abduction in both Peirce and Magnani is applicable to all perceptions. In order to appreciate the implications of such a contrast, I shall try to shed a light on Peirce's well-known view of perception in the context of animal cognition by emphasizing the double aspect of abduction as inference and instinct. Further, I shall present an interpretation of Magnani's recent studies of abduction as a sustained effort to answer how to learn abduction from animals by expanding Peircean view of perception as abduction.

> 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

This paper derives from the Lecture given at MBR012_ITALY in Honor of the 60th Birthday of Lorenzo Magnani, Chair of Conference.

W. Park (✉)
Korea Advanced Institute of Science and Technology, Daejeon, South Korea
e-mail: woosukpark@kaist.ac.kr

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_12, © Springer-Verlag Berlin Heidelberg 2014

objects, like the faculty which judges that the wolf is to be avoided and the child is to be loved. [1, pp. 30–31].

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 man's 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? [15, pp. 277–278; 24, 5.591].

From this Peircean perspective hypothesis generation is a largely instinctual and non-linguistic 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. [15, p. 286].

1 Introduction

Magnani's recent discussion of animal abduction sheds considerable light on both instinctual and inferential character of Peircean abduction. Inspired by this, I elsewhere noted some analogies and disanalogies between Avicenna's ideas on estimative faculty of animals and Peirce's and Magnani's views on animal abduction. Also, I confirmed the dividing role and function of the Beast-Machine controversy in the history of the study of animal cognition. In this paper, I propose to discuss rather extensively some of the most salient differences between Avicenna and Peirce-Magnani. Unlike estimation that only allows animals to sense what is insensible, i.e., intentions, abduction in both Peirce and Magnani is applicable to all perceptions.

In Sect. 2, I shall briefly summarize my previous discussions of some analogies and disanalogies between estimation and abduction in animals. In Sect. 3, I shall try to shed a light on Peirce's well-known view of perception in the context of animal cognition by emphasizing the double aspect of abduction as inference and instinct. In Sect. 4, I shall present an interpretation of Magnani's recent studies of abduction as a sustained effort to answer how to learn abduction from animals by expanding Peircean view of perception as abduction.

2 Analogies and Disanalogies between Estimation and Abduction in Animals[1]

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." [20, p. 131]. 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" [20, p. 131]. Fortunately, Lorenzo Magnani's recent discussion of animal abduction sheds light on both instinctual and inferential character of Peircean abduction (See [15, 17]).[2] Contrary to many commentators, who find conflicts between abduction as instinct and abduction as inference, he claims that they simply co-exist. Inspired by Peirce's and Magnani's discussions of animal abduction, elsewhere I compared Peirce's and Magnani's views of animal abduction with Avicenna's views of the estimative power of non-human animals and humans. Let me briefly summarize what I found in that comparison.

I think that we can find intriguing analogies between estimation and abduction in animals at the level of problem, diagnosis, and prognosis:

> (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.
>
> (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.
>
> (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 [21].

On the other hand, 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

[1] Section 2 is drawn from [21, 22].

[2] Magnani [15], especially Chap. 5 "Animal Abduction: From Mindless Organisms to Artifactual Mediators"

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 [22].

3 Peirce on Perception as Abduction

This third disanalogy between estimation in animals in medieval Aristotelian psychology and abduction in animals in Peirce's psychology and Magnani's cognitive science is the focal point of my discussion. In what context and for what purpose did Peirce and Magnani claim that perception is a special kind of abduction? What far-reaching implications are there in this apparently radical and controversial claim?

It is rather well-known that Peirce suggested perception as a kind of abduction. Many commentators, including Magnani, have provided us with fine expositions of Peirce's views in that regard based on ample textual ground. We will examine some such texts in due course. But, before analyzing them, it seems pertinent to locate Peirce against the background of the history of the study of animal cognition and animal souls, or more broadly the history (or pre-history) of psychology. For Peirce seems to be a figure in the transitory period between the traditional Aristotelian faculty psychology and the emerging experimental psychology.

Though rarely highlighted, there seems to be no doubt that Peirce was one of the leading American experimental psychologists in the late nineteenth and the early twentieth century. Let us find the birth date of experimental psychology in "1879", when Wilhelm Wundt established his lab at Leipzig. And let us ask what Peirce as a psychologist was doing before and after the birth of experimental psychology. Relying on Thomas C. Cadwallader's study, we may answer this question rather straightforwardly. According to Cadwallader, there were three different approaches to psychology in the mid nineteenth century: (1) the philosophical approach in Europe, which follows the Cartesian tradition of relating physiology to psychology (2) phrenology in America, which "was largely popular and had little direct input into academic psychology", and (3) the dominant psychology in America, which was "a blend of theology and philosophy" [3, p. 168]. Also, following the lead of Cadwallader, we can trace Peirce's interest in psychology back to his teen age period, for there are many interesting items in his early notebooks that deal with psychology [3, pp. 168–170]. Cadwallader also observes that after graduating Harvard in 1859 Peirce gave in his writings "an increasing focus on psychological topics" [3, p. 169]. What is important is that in this period Peirce began to criticize introspective psychology as untrustworthy in some of his famous work.[3] More importantly, Peirce criticized severely the British tradition of faculty psychology in his 1869 review of Noah Porter's book *The Human Intellect* (1868) "for failing to follow the lead of Wundt." [3, p. 171].[4]

It is tempting to dwell further on Peirce's particular achievements as an empirical psychologist. For example, we may want to uncover the Peircean heritage in psychology at Johns Hopkins even before G. Stanley Hall founded the psychology laboratory there. [3, p. 176; 8, 14]. Indeed, Cadwallader enumerates several prominent psychologists "as students and/or members of his Metaphysical Club": Jastrow, John Dewey, J. McKeen Cattell, and Christine Ladd Franklin. Or, we may want to have an overview of what Peirce wrote about topics in psychology, whether it be the Bezold-Brücke phenomenon, which Cadwallader suggests to call Bezold-Peirce-Brücke phenomenon [3, p. 172], or the problem of habit, which Peirce calls "the very market place of psychology" (7.367; [3], p. 175). But it is time to return to review how Peirce understood perception as a kind of abduction.

Perhaps the most widely cited text for Peirce's view of perception as abduction is the following:

> The third cotary[5] proposition is that abductive inference shades into perceptual judgment without any sharp line of demarcation between them; or in other words our first premises, the perceptual judgments, are to be regarded as an extreme case of abductive inferences, from which they differ in being absolutely beyond criticism. [EP 2, p. 227].

Indeed, Magnani uses this passage as the crucial evidence for the view that "Perception is abductive in itself" [15, p. 268]. And, I fully agree with Magnani for this. But what exactly do we mean by this? It is interesting to note that, in his subsequent discussion immediately following this, Magnani points out that "[i]f perceptions are abductions they are basically withdrawable, just like the scientific hypotheses abductively found". [Ibid.] Further, after discussing the semiotic and abductive character of perception with the focus on controllability, he ultimately summarizes the outcome of the discussion as follows:

> In summary, judgments in perception are fallible but indubitable abductions—we are not in any condition to psychologically conceive that they are false, as they are unconscious habits of inference [15, p. 269].

[3] Cadwallader cites "On a New List of Categories" (1867) [CP 1.545–1.559] and "Questions Concerning Certain Faculties Claimed for Man" (1868) [CP 5.213–5.263] in this regard ([3], p. 170–171).

[4] Based on Peirce's own recollection and the evidence from the large set of notes that began aroud 1865 (Ms. 1956), Cadwallader notes that "[a]s the 60s progressed, Wundt's influence began to be apparent in Peirce's writings". Also, based on a large notebook (Ms. 1156), Cadwallader reports that Peirce showed continued interest in Wundt by referring to Wundt's Physiological Psychology of 1874 at least 47 times [3, p. 171].

[5] According to Campbell, the word "cotary" is a neologism from Latin, meaning "whetstone". So, Peirce's three cotary propositions of pragmatism are supposed to sharpen the concept of pragmatism [4]. I am indebted to Lorenzo Magnani for this reference.

Again, I have no qualms with Magnani's masterly move uncovering such insightful implications of the thesis that perception is abductive in itself. However, I think it profitable to consider some other perspectives too for understanding this remarkable thesis. Anyway, Piece himself was exploiting his third cotary proposition in the context of explaining why pragmatism is nothing but the logic of abduction. At least, we should fathom how psychologists, who have been struggling with the problems of perception, would find the view that perception is a kind of abduction.

Tiercelin understands what Peirce was doing in his discussion of perception as abduction as providing us with "true connecting links between abductions and perceptions, midway between a *seeing* and a *thinking*" [30, p. 393]. Further, she finds Peirce as illustrating such connecting links by three different kinds of experiences: (1) optical illusions (2) phenomena that "involve both our constitution as a natural tendency to *interpret* and some *intentional* characteristics of the objects themselves", and (3) cases where "we can repeat the sense of a conversation but we are often quite mistaken as to what words were uttered" ([26], pp. 228–229; [25], pp. 393–394)

Perhaps, for my present purpose, Tiercelin's reconstruction of the historical and/or theoretical background, against which Peirce was presenting his views of perception as abduction, is even more pertinent. For, roughly speaking, she claims that Peirce's stance on perception evolved from that of emphasizing the inferential character of perception to an "immediate theory of perception". Since it will become extremely important for my interpretation of Peirce's view of perception as abduction, I would like to quote rather extensively how Tiercelin contrasted Peirce's early and later views of perception. Tiercelin finds in Peirce's earliest writings (1865–1868) views reminiscent of Berkeley and Helmholtz:

> In the three articles of 1868, Peirce interprets perception, from top to bottom, as an inferential and hypothetical process: he refuses to admit first premises such as sense data, impressions or intuitions (i.e., cognitions not determined by previous cognitions). Not only is space known by inference (and not immediately seen) (W 2: 197), but it is hard to distinguish dream from reality (W 2: 196) [25, p. 390].

According to her, many of Peirce's earlier views have not changed, when he presented his views on perception as abduction in 1903. "[S]ince the 1880s at least", she notes, however, Peirce adopted new ways of explaining the relations between thought and reality:

> What is most important now is to underline the external pressure, forcefulness, 'percussivity' (*NEM* 318), or reactive role of Secondness (*CP* 7.620), the 'dumbness' (*CP* 7.622) of the percept, which makes 'no professions of any kind' (*CP* 7.619) but 'obtrudes itself upon me in its entirety' (*CP* 7.624, 7.643) and the uncontrollable (*EP* 2: 191), 'unreasonable,' but for that very reason, acritical situation in which the perceiver finds himself (*CP* 7.643; *EP* 2; 191). Paradoxically, but to a certain extent only (Hookway 1986), whereas Peirce had begun as an anti-foundationalist, he now seems ready to view the perceptual judgment as (although, it is true, only *de facto*), 'the starting-point or first premise of all critical and controlled thinking' (*CP* 5.181). 'Logical criticism cannot go

> beyond *perceptual facts* which are the first judgments which we make concerning percepts' (*CP* 7.198). [25, p. 391] (Italics are Peirce's).

It is intriguing to note at this stage that the rivalry between the inferential theory of perception and the immediate theory of perception is not over. We can witness this fact by simply referring to recent articles such as [18], where he contrasts what he calls "the constructivist and ecological theories". Another and a more intriguing is Peirce's transition from an inferential theory to an immediate theory of perception. For one might simply assume Peircean view of perception as abduction to be a kind of inferential theory of perception. Josephson seems to be a good example:

> There is a long tradition of belief in philosophy and psychology that perception relies on some form of inference [2, 5, 9, 13, 28]. But this form of inference has been typically thought of as some form of deduction, or simple recognition, or feature-based classification, not as abduction. In recent times researchers have occasionally proposed that perception, or at least language understanding, involves some form of abduction or explanation-based inference [5, pp. 88, 104] [10, 11]. Peirce actually says in one place, "Abductive inference shades into perceptual judgment without any sharp line of demarcation between them" [23, p. 304]. [12, p. 238]

Insofar as one is preoccupied with the idea that abduction is a kind of inference, it would be extremely hard, if not impossible, to understand what is meant by "Peirce's transition from an inferential theory to an immediate theory of perception".

The key for solving this apparent puzzle can be found in the double aspect of abduction, i.e., abduction as instinct and abduction as inference. I suggest a hypothesis regarding the evolution of Peirce's views on abduction:

> (H1) Peirce's transition from an inferential theory to an immediate theory of perception parallels his transition from the more inferential view of abduction to the more instinctual view of abduction.

Further, I suggest another hypothesis regarding what caused these parallel transitions in Peirce:

> (H2) What Peirce learned from animals sometime in 1870s was the crucial factor in Peirce's changing mind about perception and abduction.

Needless to say, what Peirce learned from animals must be that perception is a special kind of abduction. So, my suggestions are amount to this:

> (H3) Peirce learned from animals that perception is a kind of abduction.

Again, it is needless to emphasize that I learned all these possibilities from Magnani's writings on animal abduction. So, let us turn to what Magnani has to say about animal perception and abduction.

4 Magnani on Animal Abduction

Even though Magnani never explicitly claims (H1) (H2) or (H3), I believe, they must be already implied by his ideas on animal abduction. Indeed, he characterizes the Chap. 5 of [15], which is entitled as "Animal Abduction. From Mindless Organisms to Artifactual Mediators" as the key for the entire book:

> The resulting idea that abduction is partly explicable as a biological instinctual phenomenon and partly as a more or less "logical" operation related to "plastic" cognitive endowments of all organisms naturally leads to the remaining sections (starting from Sect. 5.3) of this key chapter [15, p. 267].
>
> Further, he presents the aim of this key chapter quite explicitly:
>
> The present chapter aims at illustrating how Peircean emphasis on the role of instincts in abduction provides a deep philosophical framework which in turn supplies an anticipatory and integrated introduction to the problem of animal hypothetical cognition [15, p. 283].

So, it may not be too farfetched to portray Magnani's approach to perception as abduction as that of a truly multidisciplinary cognitive scientist's. Also, it must be worthwhile to emphasize the fact that in dealing with the issue of "perception as abduction" Magnani has in mind "animal perception as abduction", whether it be human or nonhuman. And this reminder can be supported by his strong feeling against modern philosophy:

> Sometimes philosophy has anthropocentrically condemned itself to partial results when reflecting upon human cognition because it lacked in appreciation of the more "animal-like" aspects of thinking and feeling, which are certainly in operation and are greatly important in human behavior. [15, p. 283; see also 16, p. 225].

Furthermore, Magnani's own position regarding perception does fit quite well with our characterization of Peirce's transition from inferential theory to an immediate theory of perception. After having confirmed the strong tie between perception and reification, he appeals to Raftopoulos's recent assessment of Fodor-Churchland controversy:

> in humans perception (at least in the visual case) is not strictly modular, like [6] argued, that is, it is not encapsulated, hardwired, domain-specific, and mandatory. Neither is it wholly abductively "penetrable" by higher cognitive states (like desires, beliefs, expectations, etc.), by means of top-down pathways in the brain and by changes in its wiring through perceptual learning, as stressed by Churchland (1988) [15, p. 301; 27].

Magnani believes, even if we allow "a substantial amount of information which is theory-neutral" and "a certain degree of theory-ladenness" in perceptual learning, "this fact does not jeopardize the assumption concerning the basic cognitive impenetrability of perception":

> in sum, perception is informationally "semi-encapsulated", and also semi-hardwired, but, despite its top-down [sic] character, it is not insulated from "knowledge" [15, p. 301].

So, if I am on the right track, Magnani would not only agree with (H1), (H2), and (H3) but also agree with Peirce's views on animal perception and abduction embedded in them. As was hinted at above, however, Magnani wants to go beyond Peirce. But, exactly how could he do it?

4.1 Toward Abductive Robots

The question as to exactly how to go beyond Peirce in matters with animal perception and abduction is nothing other than asking exactly how to learn abduction from animals. Even if Magnani never explicitly formulates the issue that way, throughout the chapter on animal abduction, and in fact throughout the entire volume of *Abductive Cognition*, he is dealing with that very question. The minimum ground for such an interpretation may be found even from the mottos quoted at the beginning. There Peirce is merely raising a rhetorical question: "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?" [24, 5.591; 15, pp. 277–278] On the other hand, Magnani was explicitly affirming the existence of abductive instinct in humans: "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" [15, p. 286].

But, even if we agree with Magnani for allowing abductive instinct in humans, shouldn't we concede our inferiority to nonhuman animals for the sharpness of abductive instinct? Unlike the easiness of invoking clear examples of abductive instincts in nonhuman animals, it seems rather difficult, if not impossible to cite one in the case of humans. Where are we going to find some uncontroversial example of abductive instinct in humans? Giving up this hopeless search, shouldn't we endeavor to learn from animals how to be better abducers in any specified domain?

Now, I would like to approach in stepwise fashion. First, I would like to discuss a weak or passive sense of learning abduction from animals. There is an analogy between the problem of learning from animals how to sharpen abductive instinct and the problem of learning from experts or masters of a certain field that requires extremely sophisticated skills how to sharpen abductive instinct. There is one obvious disanalogy between these two cases, of course. Unlike the entirely inborn character of the former, the latter may involve some degree of learnability. In playing board games like chess or Baduk (Weichi, Go), we novices may not only mimic masters' moves but also try to make their intuitions for selecting and focusing on a few most promising moves our own.[6] As long as the masters' intuitions themselves are not entirely inborn but to a certain extent learnt, I think,

[6] Gobet and Chassy promise to show the key features of expert intuition in chess. Their theory "explains the rapid onset of intuition and its perceptual nature, provides mechanisms for learning, incorporates processes showing how perception is linked to action and emotion, and how experts capture the entirety of a situation. [7, p. 151].

there is more hope for solving the latter problem. Be that as it may, I think, there are enough interesting similarities between the two problems. In both problems, we do concede that we ordinary humans are inferior to nonhuman animals or some experts in a specified domain. To that extent, in both problems we may be helped by abductive logic programming. Interestingly, Magnani also considers all these issues and strategies. For example, he refers to Shanahan's abductive account of robotic vision where he interprets processes "by selecting through attention, imagery, and semantic processing" as "essentially abductive" [15, p. 302; 29]. Also, he discusses the problem of expert intuition at several places in his chapter on animal abduction.

4.2 Perceiving Affordances

Secondly, I would like to discuss a strong or active sense of learning abduction from animals. I do interpret Magnani's ideas on perceiving affordance in human and nonhuman animals as an answer to the problem of how to learn abduction from animals in this sense. As far as the problem of perceiving affordances is concerned, we don't have to confess our inferiority to nonhuman animals. For, it is we humans who have perceived affordances in some highly creative ways. However, we cannot easily claim our superiority over nonhuman animals either. For, it is roaches not humans that turn out to demonstrate better ability for survival, which may imply superiority in perceiving affordances. In a word, I think we may safely and more profitably forget the issue of inferiority or superiority. Let it suffice to say that we humans, unlike nonhumans animals, seem to have very unique abductive instinct displayed by our perceiving affordances.

Magnani would be happy with my interpretation, for he himself claims that "cognitive niche construction can be considered as one of the most distinctive traits of human cognition" [15, p. 331]. According to Magnani, both human and nonhuman animals are chance seekers, and thereby ecological engineers. They "do not simply live their environment, but actively shape and change it looking for suitable chances" [15, p. 319]. Further, "in doing so, they construct cognitive niches" [Ibid.]. Then, in chance seeking ecological engineering in general, and in cognitive niche construction in particular, what exactly does differentiate humans from nonhuman animals?

In order to answer this question, we need to understand in what respects Magnani extends or goes beyond Gibson's notion of affordance. In principle, it should not be too difficult, because Magnani himself indicates explicitly or implicitly some such respects of his own innovation. Magnani takes Gibson's notion of affordance "as what the environment offers, provides, or furnishes" as his point of departure. He also notes that Gibson's further definitions of "affordance as (1) opportunities for action (2) the values and meanings of things which can be directly perceived (3) ecological facts (4) implying the mutuality of perceiver and environment" may contribute to avoiding possible misunderstanding

[15, p. 333]. Given this Gibsonian ecological perspective, Magnani appropriates some further extensions or modifications by recent scholars in order to establish his own extended framework for the notion of affordance. It is simply beyond my ability to do justice to all elements of Magnani's extended framework for affordances. Let me just note one issue in which Magnani shows enormous interest, i.e., Gibsonian direct perception.

Magnani takes Donald Norman's ambitious project of reconciling constructivist and ecological approaches to perception seriously [15, 6.4.3, pp. 343ff]. Above all, Magnani notes that Norman "modifies the original Gibsonian notion of affordance also involving mental/internal processing" [15, p. 337] based on a text, where Norman writes:

> I believe that affordances result from the mental interpretation of things, based on our past knowledge and experience applied to our perception of the things about us". [18, p. 14].

If Norman is right, we may safely infer, as Magnani does, that *pace* Gibson, "affordances depend on the organism's experience, learning, and full cognitive abilities" [15, p. 337]. Both Norman and Magnani are evidencing these ideas by formidable array of recent results in cognitive experimental psychology and neuroscience [15, p. 341].

Now, given this extended framework for that extends and modifies some aspects of the original Gibsonian notion of affordances, what exactly is Magnani's contribution? In some sense, this is an unnecessary stupid question, for everybody already knows the correct answer. By his expertise on abduction, and in particular his Peircean thesis of perception as abduction, Magnani contributes enormously to deepen our understanding of some truly big issues, such as how to reconcile constructivist and ecological theories of perception. So, my question aspires to understand more specifically how the Peirce-Magnani view of perception as abduction contributes in that regard. Let us suppose that the original Gibsonian notion of affordance has been extended and modified a la Norman. Would Magnani claim that such an extension or modification is impossible without abductive activities of organisms? Or, would he claim that such an extension or modification is still incomplete without abduction?

> Be that as it may, the big picture Magnani presents is this:
>
> Organisms have at their disposal a standard endowment of affordances (for instance through their hardwired sensory system), but at the same time they can extend and modify the scope of what can afford them through the suitable cognitive abductive skills. [15, p. 348].

If we probe the question as to what exactly are involved in organisms' employment of cognitive abductive skills, Magnani would respond roughly as the following lines:

> in sum organism already have affordances available because of their instinctive gifts, but also they can dynamically abductively "extract" natural affordances through "affecting" and modifying perception (which becomes semi-encapsulated). Finally, organisms can also "create" affordances by building artifacts and cognitive niches. [15, p. 346].

There are several points that become clear from this quote, I think. First, in addition to the original Gibsonian framework for affordances, there is room for organisms to participate in perceiving affordances (in the broad sense). Secondly, abductive skills are performed by organisms in perceiving affordances. Thirdly, in such abductively perceiving affordances, perception and action are inseparably intertwined. Finally, organisms can even create affordance by abduction. Except for the first point, I think, all these seem to be due to Magnani.

5 Concluding Remarks

One might think that creating affordances by abduction is much more remarkable than perceiving affordances by abduction. Of course, there is grain of truth in such a feeling or a view. However, it seems to me that the latter is no less remarkable than the former. For, in order to create affordance by abduction, it seems necessary to perceive affordance by abduction first. One might object by saying that that very perceiving affordance by abduction required for creating affordances by abduction is remarkable, because it involves perceiving, so to say, some non-existent thing. I do not deny that perceiving some non-existent thing is truly remarkable. But the problem is that, in some sense, perceiving some non-existent thing is just a common feature of all cases of perceiving affordances. At least, affordances are not sensible by any external sense. Whether they are dispositional properties or relational properties, we seem to perceive some non-existent thing in perceiving affordances.

This observation reminds us Avicenna's lamb that sensed the intention of the harm in the wolf by the estimative faculty as an internal sense. Apparently, the intention of the harm in the wolf is by no means something sensible by any of the external senses. So, insofar as there is no internal sense, if the lamb perceives the intention of the harm in the wolf, the lamb is perceiving something non-existent. The situation is not that different in Peirce's poor chicken, which perceives the edibility of corns. Edibility is obviously not sensible by any external sense, thereby being something non-existent. Interestingly, I found another text referring to poor chicken in Peirce's writings, which again involves perceiving something non-existent but not edibility:

> That space is not immediately perceived is now universally admitted; and a mediate cognition is what is called an inference, and is subject to the criticism of logic. But what are we to say to the fact of every chicken as soon as it is hatched solving a problem whose data are of a complexity sufficient to try the greatest mathematical powers? It would be insane to deny that the tendency to light upon the conception of space is inborn in the mind of the chicken and of every animal. The same thing is equally true of time. [25, W 3, p. 317][7]

[7] "The Order of Nature" is the title of the article, from which this passage has been excerpted. It was published in Popular Science Monthly 13 (June 1878), 203–217.

As I noted above, estimation in medieval Aristotelian psychology was peculiar to perception of intentions. On the other hand, Peirce and Magnani seem to interpret any kind of perception as abduction. To those who find the analogies and disanalogies between estimation and abduction in animals important, this is another interesting contrast. Whether it be hasty or not, I do believe that Peirce's and Magnani's generalization is truly insightful. At the same time, I wonder whether it could be possible for them to make such an abductive move without the special cases of perceiving non-existent, insensible something, such as the intention of harm in the wolf or the edibility of corns in the newly hatched chicken. Herein lies a hint for how to learn abduction from animals.

Acknowledgments I am indebted to many participants of MBR 12 conference for their incisive questions and criticisms. As always, John Woods, Claudio Pizzi, and Lorenzo Magnani gave me unmistakable moral support.

References

1. Avicenna, Rahman, F.: Avicenna's Psychology. An English Translation of Kitãb Al-Najt, Book II, Chapter VI with Historico-Philosophical Notes and Textual Improvements on the Cairo Edition. Oxford University Press, London (1952)
2. Bruner, J.S.: On perceptual readiness. Psychol. Rev. **64**(2), 123–152 (1957)
3. Cadwallader, T.C.: Peirce as an experimental psychologist, transactions of the Charles S. Peirce Soc. **11**, 167–186 (1975)
4. Campbell, P.L.: Peirce, Pragmatism, and The Right Way of Thinking. Sandia National Laboratories, Albuquerque (2011)
5. Fodor, J.: The Modularity of Mind. MIT Press, Cambridge (1983)
6. Fodor, J.: Observation reconsidered. Philos. Sci. **51**, 23–43 (1984)
7. Gobet, F., Chassy, P.: Expertise and intuition: a tale of three theories. Mind. Mach. **19**, 151–180 (2009)
8. Green, C.D.: Johns Hopkins's first professorship in philosophy: a critical pivot point in the history of American psychology. Am. J. Psychol. **120**(2), 303–323 (2007)
9. Gregory, R.L.: Perception as hypotheses. In: Gregory, R.L. (ed.) The Oxford Companion to the Mind, pp. 608–611. Oxford University Press, New York (1987)
10. Hobbs, J.R., Stickel, M., Appelt, D., Martin, P.: Interpretation as abduction. Artif. Intell. J. **63**(1–2), 69–142 (1993)
11. Josephson, J.R.: Explanation and induction, Ph. D. diss. Department of Philosophy, The Ohio State University, Columbus (1982)
12. Josephson, J., Josephson, S. (eds.): Abductive Inference. Cambridge University Press, New York (1982)
13. Kant, I.: Critiques of Pure Reason (Norman Kemp Smith, Trans.), St. Martins Press, New York (1787, 1968)
14. Leary, D.E.: Between Peirce (1878) and James (1898): G. Stanley Hall, the origins of pragmatism, and the history of psychology. J. Hist. Behav. Sci. **45**(1), 5–20 (2009)
15. Magnani, L.: Abductive Cognition. The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Berlin (2009)
16. Magnani, L.: Mindless abduction: from animal guesses to artifactual mediators. In: Bergman, M., Paavola, S., Pietarinen, A.-V., Rydenfelt, H. (eds.) Ideas in Action: Proceedings of the Applying Peirce Conference, pp. 224–238. Nordic Pragmatism Network, Helsinki (2010)

17. Magnani, L.: Is Instinct Rational? Are Animals Intelligent?: An Abductive Account. In Carlson, L., Hoelscher, C., Shipley, T.F. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Cognitive Science Society, Austin, pp. 150–155 (2011)
18. Norman, J.: The Psychology of Everyday Things. Basic Books, New York (1988)
19. Norman, J.: Two visual systems and two theories of perception: an attempt to reconcile the constructivist and ecological approaches. Behav. Brain Sci. **25**, 73–144 (2002)
20. Paavola, S.: Peircean Abduction: Instinct or Inference? Semiotica **153-1/4**, 131–154 (2005)
21. Park, W.: Abduction and estimation in animals. Found. Sci. **17**, 321–337 (2012a)
22. Park, W.: On animal cognition: before and after the beast-machine controversy. In: Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science, Western and Eastern Studies, Sapere 2, pp. 53–74. Springer, Heidelberg (2012b)
23. Peirce, C.S.: Perceptual judgments. In Buchler, J. (ed.) Philosophical Writings of Peirce, pp. 302–305. Dover, New York (1955) (1902)
24. Peirce, C. S.: Collected Papers, 8 vols., Hartshorne,C., Weiss, P. (vols. I–VI), and Burks, A. W. (vols. VII–VIII) (eds.) Harvard University Press [Abbreviated as *CP*], Cambridge (1931–1958)
25. Peirce, C. S.: Writings of Charles S. Peirce: A Chronological Edition. In Kloesel, C. J. W. (ed.) Bloomington, vol. 3. Indiana University Press [Abbreviated as *W*], Indiana (1986)
26. Peirce, C. S.: The Essential Peirce: Selected Philosophical Writings, Vol. 2, ed. N. Houser and C. Kloesel, Bloomington and Indianapolis: Indiana University Press. [Abbreviated as *EP*] (1998)
27. Raftopoulos, A.: Is perception informationally encapsulated? The issue of the theory-landenness of perception. Cogn. Sci. **25**, 423–251 (2001b)
28. Rock, I.: The Logic of Perception, MIT Press, Cambridge, MA (1983)
29. Shanahan, M.: Perception as abduction: turning sensor data into meaningful representation. Cogn. Sci. **29**, 103–134 (2005)
30. Tiercelin, C.: Abduction and the semiotics of perception. Semiotica **153**, 389–412 (2005)

Abduction and Model Based Reasoning in Plato's *Meno*

Priyedarshi Jetli

Abstract In the *elenchus* of *Meno*, Socrates employs simultaneous algorithmic and abductive visual model-based reasoning. Even though the algorithmic method would quickly provide the answer, Socrates' purpose is to make the slave boy recollect the Form of Diagonal. Recollection itself is abductive discovery and hypothesis generation. Contrary to standard interpretation true opinion rather than knowledge is recollected. For knowledge, a tether, an account or justification is required that cannot be recollected. Rather it involves abduction–deduction–induction chains of reasoning. The algorithm method is also deficient because whereas the squaring algorithm is easily grasped and employed by the slave boy, the inverse square rooting algorithm is not available to him and would be extremely difficult for him to grasp for he has not been educated in mathematics. The visual abductive model which involves counting as well as seeing is hence essential for the boy to acquire knowledge of a simple geometric proposition.

HAIL to thee blithe abduction!

Deduction or Algorithm thou never wert—
That from discovery or recollection
True opinions from your soul
In profuse generation of visual model-based reasoning art[1]
Adapted from Percy Bysshe Shelley, 'To a Skylark' (1820)

[1] Some of the words of Shelley's poem 'To a Skylark' [26] have been changed, the spirit is maintained.

P. Jetli (✉)
Department of Philosophy, University of Mumbai, Mumbai, India
e-mail: pjetli@gmail.com

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_13, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

Plato (c.427–c.347 BCE) is recognized as the master of deduction. His use of induction and abduction is most often overlooked but it is methodologically embedded in most of his deductions. There is an abundance of induction and abduction in Plato including analogical reasoning and model-based abduction. Most of Plato's compact reasoning provides paradigm examples of Magnani's 'abduction–deduction–induction cycle' [12, p. 77]. Olsen states: 'In general Plato presents puzzles, problems, and incomplete analysis, from which the reader may infer (abduct) the solutions (or adequate hypotheses)' [19, p. 86].

Plato's *elenchus* in *Meno* is an abduction–algorithm–induction cycle. The purpose of this part of the dialogue is to conclude that knowledge is recollection, and recollection is achieved through the dialectic. Dialectic is considered to be abduction by Olsen [19, p. 88].

2 Stages of the Socratic *Elenchus* with the Slave Boy

The *elenchus* is from *82b* to *85b* [21]:

I. Socrates draws a square of length two and asks the boy what the area is.

   - The boy understands that the area is four.

II. Socrates asks the boy what is the length of the square with double this area of four?

   - The boy responds that it would be the double of two, that is, four.

III. Socrates demonstrates by sketching a square with side four, and drawing boxes inside of one square unit each, what the area of the square of length four would be.

   - The boy understands that this area would be 16, four times the area of the original square.

IV. With the help of Socrates' questioning and prodding, the boy comes to realize that the length of the side of the square with area eight will be less than four but greater than two.

V. Socrates asks again, what will the length of this side be?

   - The boy responds quickly again: "three".

VI. Socrates demonstrates to the boy and the boy understands that the area of the square with side of length three will be nine.

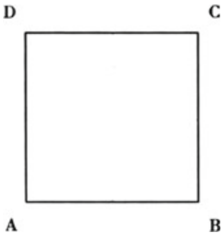
VII. The boy now admits that he is in a state of confusion.

VIII. Now, Socrates draws a diagonal of the square and constructs a square on the diagonal.

– So the boy is able to understand that the square built on the diagonal of the original square of length two will have the area double of the area of the original square, i.e. eight.

I. Abductive (Visual)–Algorithmic Model

Socrates draws a square of side length two and asks: 'you know that a square is a figure like this?' [21, *82b*][2]

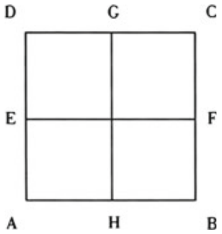


(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, accessed on 11 December 2012.)

Even though Socrates draws a nominal square of side length two, he asks whether the boy knows that a square is a figure of this type. Hence, Socrates appeals to the boy to recollect the Form of Square of which this particular square is an instantiation and simultaneously wants the boy to generalize square from this particular visual square so that the properties of the abstract square can be applied to this concrete instantiation. The visual square then represents an abductive visual model described by Magnani: '[…] as Peirce noted, abduction plays a role even in relatively simple phenomena. *Visual abduction* […] occurs when hypotheses are instantly derived from a stored series of previous experiences' [12, p. 42]. The purpose is the acquisition of knowledge by recollection, and once the Form of Square is recollected it is applied to know that what is seen is a square because it is an instantiation of the Form Square. Knowledge of Forms is knowledge of universal definitions so that when the boy grasps the definition of 'square' then he can easily answer Socrates' next question: 'It has all four sides equal?' Since having four sides equal is a necessary condition in the *definiens* of the definition of 'square', the boy responds 'Yes' [21, *82c1–2*]. 'It' refers simultaneously to the visual square that has been drawn with side length equal to two as well as to the abstract square of any side length. This is neither universal instantiation nor universal generalization. It is not a deductive model, but a combination of a visual abduction, an abductive recollection and an algorithmic calculation.

[2] This diagram is obviously drawn by Socrates but is not shown in the dialogue, but only the square at the next stage with the bisectors is shown.

Socrates immediately draws the bisectors and makes the boy admit that EF = GH.



[21, *82c*]

(Square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 8, accessed on 11 December, 2012.)

Why does Socrates introduce the bisectors?

SOCRATES: Now if this side is two feet long, and this side the same, how many feet will the whole be? Put it this way, if it were two feet in this direction and only one in that, must not the area be two feet taken once?
BOY: Yes.
SOCRATES: But since it is two feet this way also, does it not become twice two feet?
BOY: Yes.
SOCRATES: And how many feet is twice two? Work it out and tell me.
BOY: Four [21, *82c–d*].

When Socrates says: 'if this side is two feet long, and this side the same, how many feet will the whole be?', the boy could take this to mean 'when a side of the square is two feet long, then what will the area of the square be?' which the boy could answer by simple calculation 'two times two equals four'. And this the boy can easily do at this point since Socrates has already made him recollect the Form and hence the definition of 'square'. But the recollection itself is abductive. Socrates wants the boy to be able to see that the bisectors create four equal one by one squares, so that he can simply count the number of squares as four, the total area of the square. This is an effective abductive visual model. Hence, the visual and algorithmic models are simultaneously used. The visual model is abductive. One model provides a reason for the other, as Moriarti states: '[…] abduction provides a logical explanation for visual interpretation […]. Abduction begins with observation—and observations are usually visual' [15, p. 181]. The connection between the algorithmic and the visual model is abductive as the algorithm provides the algebraic answer for the visual counting. The visual (counting) and the algorithmic (calculating) hence support each other.

II. **Failure of Finding Square Root Algorithm Leading to Error**

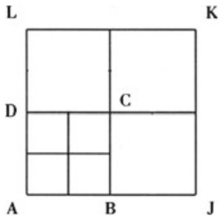
> SOCRATES: Now could one draw another figure double the size of this, but similar, that is, with all its sides equal like this one?
> BOY: Yes.
> SOCRATES: How many feet will it be?
> BOY: Eight.
> SOCRATES: Now then, try to tell me how long each of its sides will be. The present figure has a side of two feet. What will be the side of the double-sized one?
> BOY: It will be double, Socrates, obviously.
> SOCRATES: You see, Meno, that I am not teaching him anything, only asking. Now he thinks he knows the length of the side of the eight-foot square?
> MENO: Yes.
> SOCRATES: But does he?
> MENO: No [21, *82d–e*].

Socrates continues to employ the simple algorithm in getting the answer eight for the square double the area of the square with area four. The boy need look at no figure to get this answer, but simply needs to understand what 'double of' means. Now, Socrates throws a monkey's wrench, asking in stride, 'how long each side of the square double the area of the original square would be?' The slave boy, deceived in thinking that the answer will be just as automatic as the previous answer, says 'double'. If asked why the area of the square double of the given square would be eight, the boy would immediately respond because the double of four is eight. But here he cannot immediately give an algorithm. In fact if he understood the squaring algorithm he would not have given the wrong answer. In other words, the boy can employ the squaring algorithm without understanding what the algorithm is. And even if he understood it, he surely would not understand the inverse algorithm of the square root, nor would he understand that the length of the side of a square is the square root of the area. He may have understood this if Socrates had begun with drawing a square and given its area as four without telling the boy what the side of the square was, and then asked the length of the side, but then Socrates could not have accomplished the purpose of running the algorithmic method and the visual model simultaneously.

III. **Visual Model to Realize the Error**

Socrates now clearly demonstrates to the boy that the area of the square with length four, double of the side of the original square, will give us a square with area 16 and not the required eight[3]:

[3] Like the very fist square this square is not drawn by itself by Socrates in the dialogue but the completed square with the square of side three also displayed in it drawn as a composite.



[21, *83a-b*]

(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 10, accessed on 12 December 2012.)

Let us look at how Socrates comes to ‘16’:

SOCRATES: How big is it then? Won’t it be four times as big?
BOY: Of course. {visual model}
SOCRATES: And is four times the same as twice?
BOY: Of course not. {algorithmic}
SOCRATES: So doubling the side has given us not a double but a fourfold square?
BOY: True. {visual–algorithmic}
SOCRATES: And four times four are sixteen, are they not?
BOY: Yes [21, *83b–c*]. {algorithmic}

After making the boy see his error visually as the area we get is four times not double, Socrates immediately turns to purely algorithmic reasoning without bringing in the vision of the square at all. This is to emphasize the importance of the algorithmic method. It seems like that the boy will ultimately have to understand the notion of and the algorithm of square root as well.

IV. **Deductive Inference to Length being Between Two and Four**

> SOCRATES: Then how big is the side of the eight-foot square? This one has given us four times the original square, hasn’t it?
> BOY: Yes [21, *83c*].

Socrates asks two questions, but the boy answers only the second question and does not even attempt to answer the first. Why? The boy is now realizing that the answer will not come easy but he is sure of the answers found so far, so he confidently responds ‘yes’ to the second question. Neither does Socrates expect the boy to answer the first question that is why he does not persist with it in the immediately following sequence.

Socrates makes the boy realize that the length of the side of the square with area eight, double of the area of the original square with side length two will be greater than two but less than four:

SOCRATES: Good. And isn’t a square of eight feet double this and one half that?
BOY: Yes.

SOCRATES: Will it not have a side greater than this one but less than that?
BOY: I think it will.
SOCRATES: Then the side of the eight-foot figure must be longer than two feet but shorter than four?
BOY: It must [21, *83c–d*].

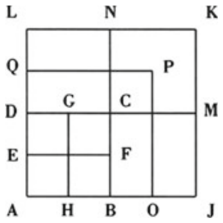
V. **Abductive Guessing**

SOCRATES: Try to say how long you think it is?
BOY: Three feet [21, *83e*].

Understanding this topological point the boy now ventures another guess, that of three. But Socrates had already anticipated this. The progression of guesses, from positing four first to positing three now after realizing that the length is less than four but greater than two, is abductive reasoning. Guessing is to be taken as a meaningful step by the slave boy towards the acquisition of knowledge, as Peirce says: '[…] every step in the development of primitive notions into modern science was in the first instance mere guess-work, or at least mere conjecture. But the stimulus to guessing, the hint of the conjecture, was derived from experience' [20, CP 2.755].[4] The experience here is the visual models being drawn by Socrates combined with algorithmic thinking. As Magnani states one of the purposes here is: 'to illustrate the relevance of the activity of guessing hypotheses, dominant in *abductive reasoning*, […]' [12, p. 2].

VI. **Visual Model Again to Demonstrate New Error**

Socrates quickly makes a three by three square in the same figure by adding the segments BO and DQ to AB and AD respectively:



(Square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 10, accessed on 12 December 2012.)

SOCRATES: If it is three feet this way and three feet that, will the whole area be three times three feet?
BOY: It looks like it. {visual}

[4] Accessed from [16, p. 218].

SOCRATES: And that is how many?
BOY: Nine. {algorithm} [21, *83e*]

This is a repeat process of when the guess of the length was four. However, this time we quickly get to nine as the area of the square with side length three, but again with both the visual model and the algorithmic model working hand in hand. Even though the boy says 'it looks like it' he is not using the visual model to calculate by counting squares but the algorithm. Since the diagram was altered by extending EF and HG to new vertices P and Q, Socrates could easily have created the nine one by one squares inside the three by three square. And then the boy could have answered visually, perhaps he sees it anyway. But Socrates wants the boy to be thinking in the stream of the algorithm. What follows about the correct length of the side of the square with area eight and the figure that displays that cannot be confirmed by counting one by one squares, as in the case with perfect squares, that is those of areas four, 16 and nine.

VII. **Reaching the State of Confusion before the Final Step to Recollection**

> SOCRATES: Whereas the square double our first square had to be how many?
> BOY: Eight.
> SOCRATES: But we haven't got the square of eight feet even from the three-foot side?
> BOY: No.
> {The emphasis on 'three' here is Socrates' hint or rather the obvious inference that the length of the side of the desired square will not be a natural number and this leaves the boy perplexed}
> SOCRATES: Then what length will give it? Try to tell us exactly. If you don't want to count it up, just show us on the diagram.
> BOY: It's no use Socrates, I just don't know. [21, *83e–84a*] {The state of confusion}

The choice that Socrates gives the boy is between counting and seeing it on the diagram. Both are part of the same process as the dialogue has proceeded so far. The counting is to be done on the diagram not independent of it and is hence part of the visual model. All through Socrates has employed simultaneous visual model and algorithmic reasoning. 'Show us' however challenges the boy to use the visual model to demonstrate the length of the side of the square we want, but the boy is unable to do it at the moment. Algorithm has been used exactly three times in 'two times two is four', 'four times four is sixteen', and 'three times three is nine'. Then why not direct the boy towards the algorithmic 'x times x is eight'? Because the square root of eight is an incommensurable number and neither Socrates nor anyone in his time understood incommensurable numbers so how could Socrates expect the boy to understand this. Nor does the boy at this point understand the converse algorithm that the square root of four is two, the square root of sixteen is four and the square root of nine is three.

SOCRATES: Observe Meno, the stage he has reached on the path of recollection. At the beginning he did not know the side of the square of eight feet. Nor indeed does he know it now, but then he thought he knew it and answered boldly, as was appropriate—he felt no perplexity. Now however he does feel perplexed. Not only does he not know the answer; he doesn't even think he knows.
MENO: Quite True.
SOCRATES: Isn't he in a better position now in relation to what he didn't know?
MENO: I admit that too.
SOCRATES: So in perplexing him and numbing him like the sting ray, have we done him any harm?
MENO: I think not.
SOCRATES: In fact we have helped him to some extent toward finding out the right answer, for now not only is he ignorant of it but he will be quite glad to look for it. Up to now, he thought he could speak well and fluently, on many occasions and before large audiences, on the subject of a square double the size of a given square, maintaining that it must have a side of double the length.
MENO: No doubt.
SOCRATES: Do you suppose then that he would have attempted to look for, or learn, what he thought he knew, though he did not, before he was thrown into perplexity, became aware of his ignorance, and felt a desire to know?
MENO: No.
SOCRATES: Then the numbing process was good for him?
MENO: I agree [21, *84a–c*].

Stage VII is the paramount feature of the Socratic *elenchus*. The student, or answerer, the boy, has reached an authentic state of confusion, which was exactly the aim of the interrogator or teacher, Socrates. This is the stage where due to the rigorous progression of the earlier stages the answerer comes to know that he does not know what he earlier claimed to know. In *Euthyphro*, Euthyphro who thinks he knows what piety is reaches the state of confusion by the end of the dialogue and instead of persisting with acquiring authentic knowledge of what piety is he simply abandons the project. In *Crito*, Crito makes a series of knowledge claims which are one by one dismissed by Socrates' interrogation leaving Crito in a state of confusion and he does not pursue any of these further as his pragmatic aim is to convince Socrates to escape from prison.

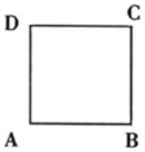
In *Meno* the boy reaches the state of confusion because now he knows that he does not know what the length of the side of the square with area eight square feet is. However, in the process of the *elenchus* he has come to know that it is neither four nor three and further that it is greater than two and less than three, so he has made remarkable progress towards authentic knowledge.

VIII. **Recollection as Discovery and Abduction**

SOCRATES: Now notice what, starting from this state of perplexity, he will discover by seeking the truth in company with me, though I simply ask him questions without teaching him. (*Socrates here rubs out the previous figures and starts again*) [21, *84c–d*].

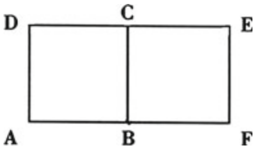
The final stage has begun where the boy will come to have knowledge of what is the length of the square with the area of eight square feet.

SOCRATES: Tell me boy, is not this our square of four feet? [ABCD.] You understand?



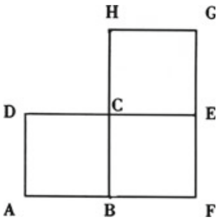
(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 12, accessed on 13 December 2012.)

BOY: Yes.
SOCRATES: Now we can add another equal to it like this? [BCEF.]



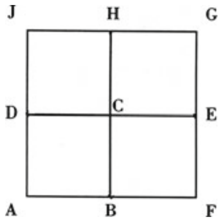
(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 12, accessed on 13 December 2012.)

BOY: Yes.
SOCRATES: And a third here, equal to each of the others? [CEGH.]



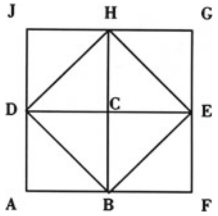
(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 12, accessed on 13 December 2012.)

BOY: Yes.
SOCRATES: And then we can fill in this one in the corner? [DCHJ.]



(Modified from square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 12, accessed on 13 December 2012.)

BOY: Yes.
SOCRATES: Then here we have four equal squares?
BOY: Yes.
SOCRATES: Then how many times the size of the first square is the whole?
BOY: Four Times.
SOCRATES: And we want one double the size. You remember?
BOY: Yes.
SOCRATES: Now does this line going from corner to corner cut each of these squares in half?
BOY: Yes [21, *84d–85a*].



(Square accessed from http://kireetjoshiarchives.com/teachers_training/good_teacher/learning_recollection.php, p. 12, accessed on 13 December 2012.)

Even though there are four lines Socrates uses the singular 'line' in order for the boy to abstract the concept of a diagonal.

SOCRATES: And these are four equal lines enclosing this area? [BEHD.]
BOY: They are.

Now, the visual model is coming on strong.

SOCRATES: Now think. How big is this area?
BOY: I don't understand.

Of course the boy does not understand as so far Socrates has dissuaded the boy from counting but now he wants him to count.

SOCRATES: Here are four squares. Has not each line cut off the inner half of each of them?
BOY: Yes.

Again the emphasis is on the visual model.

SOCRATES: And how many such halves are there in this figure? [BEHD.]
BOY: Four.

The answer here can only be gotten through seeing and counting.

SOCRATES: And how many in this one? [ABCD.]
BOY: Two.

The answer here can only be gotten through seeing and counting.

SOCRATES: And what is the relation of four to two?
BOY: Double [21, *85a*].

The answer to this question does not depend on the visual model but on a simple algorithm.

SOCRATES: How big is this figure then?
BOY: Eight feet [21, *85a–b*].

Again this answer is based on a simple algorithm as the double of four, the area of the original square, is eight. Also, both the doubles refer to areas. The four equal areas seen are double of the two equal areas seen and it is also seen that each of these two equal areas is equal to each of the four equal areas.

SOCRATES: On what base?
BOY: This one.

The visual model is now used as the boy points out to the base, the diagonal BD.

SOCRATES: The line which goes from corner to corner of the square of four feet?
BOY: Yes.

The diagonal is pointed to by Socrates while emphasizing its diagonal nature.

SOCRATES: The technical name for it is diagonal; […] it is your personal opinion that the square on the diagonal of the original square is double of its area.
BOY: That is so Socrates [21, *85b*].

This is the end of Stage VIII.

3 Recollection as Abduction

Is this recollection? It is a whimper rather than a bang. It does not seem convincing at all that the slave boy has recollected anything leave alone first the Form of Diagonal and then the knowledge that the area of the square double the area of the original square will be the square with the diagonal of the original square as its side. The boy has answered in the same expression and has not said 'eureka!' or expressed any excitement or given any indication of having made a discovery.

Has Plato cheated us? Not really! As Plato has reminded us through his rigorous toils in every dialogue, recollection does not come easy, whether it is the Form of Piety, which is never recalled by Euthyphro at the end of *Euthyphro*, or the Form of Knowledge, which Theaetetus comes close to recollecting at the end of *Theaetetus*, but falls short of it as the definition he has found so far, true belief with an account is not yet knowledge: 'So, my friend, there is such a thing as right belief together with an account, which is not yet entitled to be called knowledge' [22, *208b*]. In the *Republic* it is Plato, through Socrates, who finally provides the definition of 'justice'. Yet, one can read the dialogue many times but fail to see at what point Socrates recollects the Form of Justice. But what we find in the *Republic* in the construction of the Platonic definition of 'justice' is the finest complex of craftsmanship in philosophy. Plato did not use the examples of craft persons like carpenters, cobblers and weavers just for entertainment. He wanted to convey that philosophy as systematic thinking and theorizing requires the finest craftsmanship. Socrates as a teacher was a craftsman of unmatched skills and perseverance.

In Stage VIII we have seen a display of this craftsmanship. The dialectic has picked up pace from the earlier part where it moved rather slowly and meticulously, it is like a symphony reaching a crescendo. Socrates begins this section with a brick by brick construction of the visual square which should finally convince the boy that the area of the square with the side double of the original side will be four times the original square not two times. Not only does the boy see this for this particular square which is drawn in front of him, he is able to grasp the generalization perhaps through an abduction–deduction–induction cycle.

Next, Socrates constructs the diagonals for each of the four squares and visually demonstrates, or rather has the slave boy visually realize by looking and simple calculation that the area of the square with the diagonal as the side is the desired eight, as it is four halves of the square with area four. At this point one would hope that the boy has a flash and has discovered a geometrical truth about the area of the square of the diagonal of a square to be double of the area of the original square. Perhaps, if the boy has not actually reached the moment of discovery, the same procedure could be repeated with squares of different lengths and sooner or later the discovery will come.

Isn't this the daily procedure of so many human processes from pottery to music to tennis to mathematics? Each of us must have so many childhood memories when we had difficulty with understanding some concept such as why the square root of a positive number could be either positive or negative, and finally some flash came and we understood it and have understood it ever since. Socrates is well aware that the slave boy even if he has had a flash cannot at the moment describe it:

SOCRATES: What do you think, Meno? Has he answered with any opinions that were not his own?
MENO: No, they were all his.
SOCRATES: Yet he did not know, as we agreed a few moments ago.
MENO: True.
SOCRATES: But these opinions were somewhere in him, were they not?
MENO: Yes.
SOCRATES: So a man who does not know has in himself true opinions on a subject without having knowledge.
MENO: It would appear so.
SOCRATES: At present these opinions, being newly aroused, have a dreamlike quality. But if the same questions are put to him on many occasions and in different ways, you can see that in the end he will have a knowledge on the subject as accurate as anybody's.
MENO: Probably.
SOCRATES: This knowledge will not come from teaching but from questioning. He will recover it for himself.
MENO: Yes.
SOCRATES: And the spontaneous recovery of knowledge that is in him is recollection, isn't it?
MENO: Yes [21, *85b–d*].

Even though Socrates uses the term 'spontaneous recovery' we can term it as *discovery* in the Peircean sense since it is at least not consciously available to us before the dialectic process but only at the end of a long and arduous dialectic process. Recollection then as being 'spontaneous' or in a flash cannot be either deduced or induced but only abduced. Though the dialectic is the path to recollection, it is not a deductive or inductive inference to recollection, in fact recollection may or may not happen at the end of a dialectic process. How does recollection actually happen? How does the moment of 'eureka' actually happen? The inference for recollection is as follows:

The surprising flash of knowledge occurs.
But if there were recollection, flashes of knowledge would be a normal occurrence.
Hence, there is a reason to suspect that recollection happens.

This is an instance of abductive reasoning as described by Jaime Nubiola:

The surprising fact, C, is observed.
But if A were true, C would be a matter of course.
Hence, there is a reason to suspect that A is true [18, p. 126].

Almost all Plato scholars agree that recollection involves discovery. Scott states: 'Everyone would agree that the theory of recollection is intended to explain how philosophical and mathematical discoveries are made' [24, p. 7]. The explanation of how discoveries are made is the domain of abduction. As Polanyi states: '[…] scientific discovery cannot be achieved by explicit inference, nor can its true claims be explicitly stated. Discovery may be arrived at by the tacit powers of the mind, and its content, so far as it is indeterminate, can be only tacitly known' [23, p. 158].[5] We could not have a better modern account of Plato's knowledge as recollection. If we can reconcile the notion of discovery with recollection in the form of tacit knowing, then we can bring in Peirce's statement: 'Abduction is the process of forming an explanatory hypothesis. It is the only logical operation that introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a hypothesis' [20, CP 5.171].[6] Peirce may not have approved of recollection and might have taken discovery to be more authentic than Plato or Polanyi, but Mullins finds a nice synthesis: 'Examining tacit knowing in conjunction with Peirce's ideas about abduction provides a new and rich context within which to appreciate Polanyi's claims for tacit knowing' [16, p. 199]. And if Polanyi's tacit knowledge is a revival of Plato's knowledge as recollection, then Plato's recollection is a type of Peircean abduction.

Magnani expresses some doubt whether recollection can be considered as abduction: '[…] in order to solve a problem one must in some sense already know the answer, there is no real generation of hypotheses, only recollection of them' [12, p. 1]. I am in slight disagreement here as the notion of tacit knowledge in Plato has to be taken within the context of Platonic epistemology. The claim that knowledge is propositional, that the object of the verb 'know' is a proposition is firmly established in *Theaetetus*, where knowledge is also tentatively defined as true opinion with an account. Plato does have some notion of this in the *Meno* as well, as later Socrates clearly states: 'true opinion is as good a guide as knowledge for the purposes of acting rightly. […] So right opinion is something no less useful than knowledge' [21, *97b–c*]. Meno immediately throws a doubt as surely, he thinks, knowledge must be superior to true opinion. Socrates gives the reason for Meno's intuition: 'True opinions […] run away […] they are not worth much until you tether them by working out the reason. That process […] is recollection […]. Once they are tied down, they become knowledge, […] [21, *97e–98a*].' Hence, there is a closure within *Meno*. True opinion is good enough for practical knowledge, knowledge required for action; however, for theoretical knowledge, true opinion needs to be tied down as knowledge by providing an account.

This failure to be clear about what exactly is being recollected, true opinion or knowledge, leads to misconceptions of recollection being the one and all of Plato's epistemology such as that by Stefanson: 'Knowledge (*epistēmē*), on Platonic

[5] Accessed from [16, p. 207].

[6] Accessed from [16, p. 199].

understanding, is an entirely separate activity and a distinct possession to belief or opinion (*doxa*). […] Understanding is only possible when a man allows his *psychē* to contemplate the metaphysical realm of knowledge' [27, pp. 102–103]. However, I contend that what is recollected through what Stefanson calls 'to contemplate the metaphysical realm' is true opinion, whereas understanding has to do with giving an account.

Even famous Plato scholars like F. M. Cornford do not clearly see that what is being recollected in *Meno* is not knowledge but true opinion and though true opinion may be practical knowledge it is not theoretical knowledge: 'The *Meno* had already announced the theory of *Anamnesis*: that knowledge is acquired […] by recollection […] truths seen and known by the soul […] for seeking and learning is nothing but recollection. […]' [3, p. 2]. Is the use of 'knowledge' here ambiguous, referring either to 'true opinion' or to 'true opinion with an account' or true belief with understanding? What follows makes it clear that Cornford is not making a conflation: '[…] and if he were questioned again and again in various ways, he would end up having knowledge in place of true belief—knowledge which he would have recovered from his own soul. This knowledge must have been acquired before birth' [3, p. 3]. So, Cornford clearly believes that what is being recollected by the slave boy is not true opinion but knowledge. This is rather unfortunate as my exegesis of the passage shows just the opposite, namely, that as Plato takes pain to demonstrate later in *Meno*, knowledge requires a tether that is not there in true opinion and the tether cannot be recalled or recollected but must in a way be constructed as an account or understanding.

Norman Gulley also believes that what is recollected is 'knowledge': 'Thus the knowledge sought by Socrates […] is described by Plato as aknowledge of 'forms', […]' [7, p. 3]. Gulley goes on to make this assessment about: '[…] passage of the *Meno* […] that these ideas embody a rudimentary theory of recollected knowledge' [7, p. 9]. Gulley then discusses the distinction between true belief and knowledge. I believe he is in error when he describes the second stage of recollection as '[…] recognition that certain propositions are true, but not as yet *why* they are true. […] The level of apprehension now reached is described as "true belief" (85c, 86a)' [7, p. 13]. This has taken place as recollection of a true opinion as clearly stated:

SOCRATES: The technical name for it is diagonal; […] it is your personal opinion that the square on the diagonal of the original square is double of its area.
BOY: That is so Socrates [21, *85b*].

This is the point of recollection. And in the immediately following discussion with Meno Socrates emphasizes that what is recollected is not knowledge but true opinion: 'So a man who does not know has in himself true opinions on a subject without having knowledge' [21, *85c*]. So unless at the second stage Gulley means recollection by 'recognition' and 'apprehension' he is mistaken. However, Gulley, unlike Cornford does state that for Plato at this point the tether is required to make true opinion into knowledge: 'To become knowledge it must be "tied down" by a "chain of causal reasoning" […] the method of analysis […] to find the antecedent conditions […] a method practiced by geometers' [7, pp. 14–15]. Incidentally the geometric method of analysis seems to be abductive reasoning as it is reasoning to antecedents or hypotheses.

Magnani also sees that what is recollected is true opinion, not knowledge: 'The true opinion is given by recollection and science is the system of true opinions when related by the activity of reasoning and thereby made permanent and definitive' [12, p. 7]. Science then turns true opinion into knowledge by what Plato calls the 'tether', the account that gives proper understanding.

Plato himself is responsible for an equivocation in the use of the words 'knowledge' and 'know'. In the passage just quoted above, first Socrates says that true opinions, not knowledge, were in the boy: 'So a man who does not know has in himself true opinions without having knowledge'. 'True opinion' here is ambiguous between the tacit true opinion and the conscious true opinion after recollection. What is recollected is true opinion not knowledge: 'At present, these opinions, being newly aroused have a dreamlike quality'. Knowledge then is not recollected, true opinions are, and they can then be turned into knowledge: 'if the same questions are put to him […] in the end he will have a knowledge on the subject […]'. All of this seems to be clear but then Socrates says 'But this knowledge does not come from teaching but from questioning. He will recover it for himself.' The last sentence indicates that knowledge after all is recollection. So, if my earlier claim is correct that the boy has not at this point in the dialogue recollected the Form of Diagonal, but with repeated questioning he will eventually recollect and thereby have knowledge; then Magnani's line of 'tacit knowledge' is correct. This is further supported by the immediately following text: 'Either then he has at some point acquired the knowledge he now has, or he has always possessed it' [21, *85d*].

However, I insist that we take 'knowledge' in the larger context of Plato's epistemology, especially in the context of the definition of propositional knowledge established later in *Theaetetus*. In recollecting Forms one does not recollect propositions but abstract objects. Then one puts these abstract objects together with concrete particulars to form beliefs or opinions. Hypotheses after all are not objects but propositions. So it is not even true opinion that is recollected but a true opinion is formed immediately at the time of recollection.

Whether he recollects Forms or true opinions, he does not recollect an account or adequate justification of the opinion. An account is surely not something that can be recovered, but something that must be constructed from understanding. Hence, if the concept of 'knowledge' includes understanding then providing an account becomes essential. Hence, recollection, at the end of the process of the dialectic is necessary for knowledge but it is not sufficient as the additional condition of account is required which may or may not be available at the time of recollection.

This is a dynamic example. Let us suppose that not only the boy, but Socrates, Meno and everyone else present at the time, and for that matter anyone reading the dialogue for hundreds of years after that, including the greatest philosophers and mathematicians, all recollected the Form of Diagonal and reached the true opinion that the square which has double the area of the original square must be the square constructed on the diagonal of the original square; but none of them have knowledge, because whenever a square has a length that is a discrete number like 1, 2, 3, 4, 5, and so on, then the diagonal of the square will be an incommensurable number like $\sqrt{2}$, $\sqrt{8}$, $\sqrt{18}$, $\sqrt{32}$, $\sqrt{50}$, and so on. At the time of Socrates and Plato no one properly understood incommensurable numbers, so at least a complete justification was lacking, hence if complete justification is a necessary condition for knowledge,[7] then no one could have had knowledge of the hypothesis that is being generated through the abductive process of recollection here. So, we should not be surprised at all that despite recollection the boy is falling short of knowledge because he lacks an account and understanding.

Induction and abduction are intricately linked in the *elenchus* in *Meno* so that we cannot really say which comes first, as Nguifo states: '[…] in the ML systems, abduction cannot be used without induction, and induction needs abduction' [17, p. 49].

4 The Parallel Geometric Visual Model and the Algorithm

Even near the end of the dialectic part of the *elenchus* an opinion or what we today call a 'belief' is reached not as a whim but as one with a reasoning process backing it, but it is not yet knowledge. At the end of the long dialectic process one reaches a belief and one still has a long way to go in making this belief into knowledge.

Socrates, if he were interested only with the right answer, would have worked only with the algorithmic model. Reasoning as follows:

$$2^2 = 4$$
$$\sqrt{4} = 2$$
$$\sqrt{8} = 2\sqrt{2}$$

In this case none of the figures drawn would be required. But the whole point that Socrates wants to make about recollection as a progressive dialectical process of reaching true opinion would have been lost. Furthermore, the simultaneity of the visual model and the algorithm is required to come to realize that the square root of eight is an incommensurable number.

Elaine Landry claims that in *Meno* Plato has two supplemental methods running parallel, namely the *elenchus* and the hypothetical, or mathematical [10, pp. 1–2]. I would rather claim that the hypothetical method is embedded in the *elenchus* as the *elenchus* is the process. Landry's hypothetical method requires the parallel geometric visual model and an algorithm.

[7] Keith Lehrer and Thomas Paxson Jr. take 'S is completely justified in believing that *h*' to be one of the necessary conditions of knowing [11, p. 225].

The arithmetical solution requires neither drawing figures nor the calculation of the length of the diagonal by the Pythagorean Theorem. Algorithms are generally considered to be deductive, but are they? Once an algorithm is constructed then it can be used to deductively churn out a result by plugging in values. Compact deductive arguments, operating on a simple algorithm, sometimes may not lead to knowledge even if one immediately grasps the soundness of the argument.

The implicit induction here is that a method of counting works in determining the areas of squares. The use of geometric visual models by Plato is abductive model-based reasoning. The squares with side two, three and four are presented as geometric models. This brings upon the realization that the side of the square with area eight is an incommensurable number, the length of which can nonetheless be determined by the Pythagorean theorem so deduction comes into aid model-based reasoning. Furthermore the boy uses abductive reasoning of making closer guesses or conjectures as he moves from four to three to realizing that the answer is between two and three, closer to three. As Peirce states: 'The order in the march of succession in retroduction is from experience to hypothesis' [20, 2.755].

Meno presents a classic case of abduction as 'reasoning which starts from data and moves towards hypothesis' as Fann claims [5, p. 5]. When Socrates asks the initial question, the boy goes into search for a hypothesis but he does not quite know how to arrive at it. Through the dialectic process he finally arrives at the hypothesis that the length of a side of a square double the area of another square is the length of the diagonal of the original square as he recollects the Form of Diagonal. Recollection is discovery of sorts and this makes it abduction.

If the boy knew the diagonal hypothesis from the beginning he would have answered the question right away. But he does not have this hypothesis available to him but only arrives at it through the dialectic within *elenchus*. Socrates does not give the boy the option of pursuing the much easier algorithmic method because through the algorithmic method he would understand it arithmetically but would not understand it geometrically, which is the more complete knowledge that Plato was seeking. Once the first diagram is drawn, the boy could easily internally visualize with the aid of the simple algorithmic method. However, the advantage of the continual use of the visual model by Socrates is best explained by Mary Hegarty: 'viewing an external visualization […] can be a substitute for internally visualizing […] the availability of external visualizations relieves us of the necessity of internally visualizing […]' [8, p. 5].

5 Socrates–Slave Boy *Elenchus* as a Hybrid Algorithm—Abductive (Visual) Model

The *Meno* hence provides us with a paradigm hybrid method as Magnani says 'visual and algorithmic may be intertwined, and so, hybrid so to say' [14]. In modern algebraic geometry by understanding the algorithmic method one may well understand the geometrical hypothesis as well. Another reason why Socrates does not end matters with the algorithmic method is because this would not lead to the realization that the square root of 8 is an incommensurable number, which may be the ultimate aim of the inquiry, not so much for the slave boy but for the rest of us, at the time of Plato of course. Gabbay and Woods state: ‘abduction is the finding and engaging of a hypothesis (H) that, when combined with what one already knows (K), enables one to presumptively attain a cognitive target (T) that one could not attain via K (or a ready expansion of K) alone’ [6, p. 290]. If K is the knowledge available at the point before recollection of the diagonal including the knowledge of the extended algorithm, it would not be sufficient to realize that the length of the diagonal is an incommensurable number, yet its square is a commensurable number.

The use of perceptual models by Socrates of drawing squares on the ground and explaining to the slave boy through these visual models is what Peirce would call ‘abduction’: ‘[…] our first premises, the perceptual judgments are to be regarded as an extreme case of abductive inference […]’ [28, p. 393].

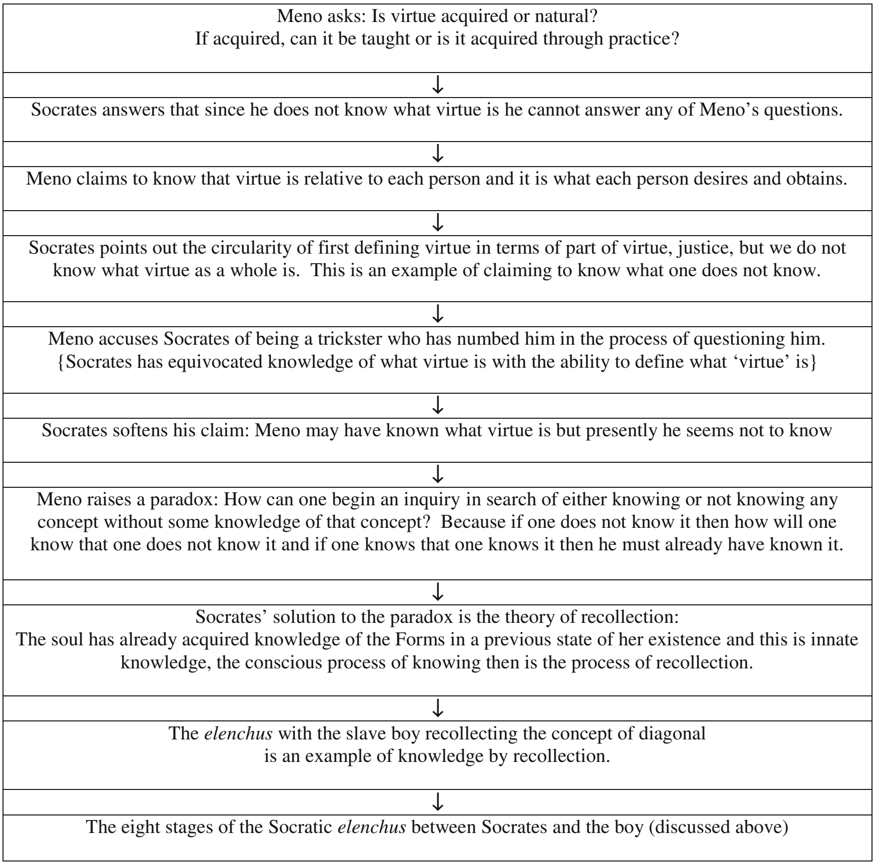
Both the algorithm and the diagrammatic model are alternative computational methods where each ‘computational model is the theory not a simple instantiation of a theory’ [9, p. 511]. The reasoning presented in the *Meno* may best be described as what Magnani calls ‘manipulative abduction’: ‘the exploitation of external logical and computational abductive–but also inductive—systems/agents to form hybrid and multimodal representations and ways of inferring in organic agents’ [13, p. 396].

6 The Surprise: No Deduction in the *Elenchus* Between Socrates and the Slave Boy

There is a great surprise after this long trek through the Socrates–slave boy *elenchus*. I started the introduction with the sentence ‘Plato is recognized as a master of deduction’. In this *elenchus* and dialectic, one of the most famous in Plato, where is deduction? There are some very simple straightforward deductive arguments in some of the eight stages, but these too are more implicit than explicit, like in deducing that the length of the side of the square we are looking for is between two and four since it is greater than two and less than four. But the spots of deduction that are present are hardly anything to write home about. If algorithm is not deduction as I have argued above, then the *elenchus* has the parallel abductive visual and guessing models and an algorithm embedded in it. Within the abduction visual and guessing as two species of abduction are embedded; and the sub-species of counting is further embedded in the visual. The counting is interwoven with induction. Within the guessing there is some minimal deduction woven in as the slave boy has realized that the length of the side must be between two and four because he knows at that point that it is greater than two and less than four; so he makes the guess of three. All of this part of the *elenchus* leads to the dialectic and recollection as a species of abduction is embedded in the dialectic. This surprise of a prime piece of Platonic reasoning that is not centered on deduction is itself a hypothesis inferred through abduction.

7 The *Meno Elenchus* in the Context of the Narrative of the Whole Dialogue

The Narrative Sequence of the *Meno*



Meno asks: Is virtue acquired or natural?
If acquired, can it be taught or is it acquired through practice?

↓

Socrates answers that since he does not know what virtue is he cannot answer any of Meno's questions.

↓

Meno claims to know that virtue is relative to each person and it is what each person desires and obtains.

↓

Socrates points out the circularity of first defining virtue in terms of part of virtue, justice, but we do not know what virtue as a whole is. This is an example of claiming to know what one does not know.

↓

Meno accuses Socrates of being a trickster who has numbed him in the process of questioning him.
{Socrates has equivocated knowledge of what virtue is with the ability to define what 'virtue' is}

↓

Socrates softens his claim: Meno may have known what virtue is but presently he seems not to know

↓

Meno raises a paradox: How can one begin an inquiry in search of either knowing or not knowing any concept without some knowledge of that concept? Because if one does not know it then how will one know that one does not know it and if one knows that one knows it then he must already have known it.

↓

Socrates' solution to the paradox is the theory of recollection:
The soul has already acquired knowledge of the Forms in a previous state of her existence and this is innate knowledge, the conscious process of knowing then is the process of recollection.

↓

The *elenchus* with the slave boy recollecting the concept of diagonal
is an example of knowledge by recollection.

↓

The eight stages of the Socratic *elenchus* between Socrates and the boy (discussed above)

↓
The Boy acquired knowledge by recollection of the Form of a diagonal through the *elenchus.* (stage VIII)
↓
Neither true opinion nor knowledge can be taught but they are acquired through questioning, the *elenchus*.
↓
If virtue were a form it would be known through the *elenchus*, but knowledge of virtue is not that.
↓
If virtue is knowledge it is teachable.
↓
Virtue is not knowledge but wisdom, that is, it is the ability to know and not knowledge itself.
↓
If virtue were teachable who would be the teachers of virtue? Would it be the Sophists?
↓
Anytus who now walks into the conversation protests that the Sophists cannot be teachers.
↓
Neither the Sophists nor those who possess and display virtue are teachers of virtue.
↓
Who then can teach virtue? No one. So, there are neither teachers nor students of virtue.
↓
True opinion and knowledge are not natural. So they are acquired.
↓
Since virtue is not teachable it is not knowledge. Yet virtue is not natural but acquired.
↓
Virtue, like true opinion, is acquired and can be practiced without being taught or known

Taking the whole narrative into account the starting point is the question: is virtue teachable? The conclusion is that virtue is not teachable and if knowledge is teachable then virtue is not knowledge. This is essentially a deductive argument. Embedded in this main argument is the *elenchus* of Socrates and the boy, which is essentially an abductive–algorithmic argument. Hence the *elenchus* in the context of the entire dialogue is embedded in an overarching deductive argument for why virtue is not teachable.

8 Conclusion

One just does not want to leave *Meno* as there is so much there. We have seen varieties of abductive reasoning in *Meno elenchus* with the slave boy, including diagrammatic, visual, guessing, and discovery (which is recollection, the heart of the narrative). If we consider model-based reasoning to be ‘the consideration and manipulation of various kinds of representations' [12, p. 45], then the algorithm and the geometric visual model are simultaneous model-based reasoning processes embedded in the *elenchus*. Plato was actually extremely complex, extremely sophisticated and way ahead of his times as any attempt to interpret him, including mine probably does not come close to what he really had in mind. Nonetheless, it would only be in Platonic spirit to entertain all interpretations but at the same time try to establish the viability of each interpretation.

Gerald Boter argues that the lines to be drawn in the beginning inside the square are the diagonals and not the transversals so that everyone who follows the literal reading of the instructions given by Socrates is mistaken. Boter's reason is that Socrates wants the boy to determine the area of the square by calculating and not by counting, in which the transversals help by dividing the square up into four one by one squares; and in the end the answer to the desired question is the diagonal so Socrates wants to draw it first [2; 25, p. 220]. This would be a viable interpretation only if Plato meant to use only one method here or thought that there was only one method. I have on the contrary shown that Socrates requires both the algorithmic method as well as the visual geometric method in order to make the full use of the dialectic towards recollection.

Boter's conjecture then is motivated by the myopic view of Plato that he used only deduction as a method, whereas what we have seen is that the dialectic is not possible without abduction as the final step of recollection itself is a stage of discovery and thereby abductive and not deductive. In fact deduction plays a marginal role or no role at all in the entire *elenchus* of *Meno*. The reason why Socrates chose the slave boy instead of a regular school boy with instructions in geometry; is that Socrates was neither trying to show off his knowledge in geometry as there were better geometers than him as the audience of the dialogue, nor was Socrates after any kind of theorem or proof. And a slave boy with no traditional schooling was not likely to either think of or be interested in a deductive proof, nor would he have understood it.

The purpose of the early point of the dialectic is for the slave boy to realize that he made an error and also to understand why he made the error. The visual method of counting is an effective pedagogical device to realize the error and then with the algorithmic 'and four times four is sixteen', and so on, helps the boy understand his error. First, in the visual square the boy sees that the square with the side four would have four squares at the bottom and hence four squares at the top, but once he simultaneously calculates this algorithmically he does not actually need to count all sixteen. This is why the two methods are purposefully used simultaneously.

The three stages of the dialectic are best captured by Bluck:

> First, Socrates dispels the slave's false supposition that a square twice the size of the original will have sides twice the length […] and his next incorrect answer, 'three feet', is treated in the same way. This elenchic or refutative procedure has a positive aspect: […] and aid towards the recollection of the correct answer […] a very important stage in the recollection process. […] The second stage is the 'stirring up' of latent true opinions, and the third is the conversion of these into knowledge [1, p. 15].

Bluck makes it clear that in order to accomplish the first stage effectively Plato requires the geometrical visual model and the algorithmic method alone would not have produced this:

> […] recollection can be aided by careful questioning and perhaps also by sense-experience, and that it is a process not a sudden jump […] recognition that neither a two- nor three- nor four-foot line will give a square of eight square feet […] lead(s) to a discovery of the correct solution […] the method of 'stirring up' true opinions on the question at issue is not radically different from the method of eliminating false opinions […] [1, pp. 16–17].

It is gratifying to find support in a scholar like Bluck of almost everything I have maintained in my discussion of Plato. The most important gratification is that what is being recollected are true opinions and not knowledge, as knowledge according to Bluck is the third stage, the tether stage. Bluck also makes an important point earlier that since the *Meno* is before the *Phaedo* and the theory of Forms does not begin to be formed until the *Phaedo*, it would not be appropriate to claim that the boy or Meno, in the dialogue, would recollect Forms [1, p. 6], and if knowledge is to be of Forms then they cannot recollect knowledge; however they could well recollect true opinions.

Finally, *86d-87c Meno* is more crucial in resolving the paradox of virtue and knowledge and this also involves the geometrical method. Mark Faller claims that the main purpose of Plato's analogy, in this passage, of the geometric problem of inscribing a triangle in a given circle and in determining whether virtue is knowledge, is not simply to show that virtue cannot be knowledge but to point out the logical form of analogical reasoning. Plato takes care to demonstrate the clear formal isomorphism between the geometric problem and the *Meno* paradox of virtue and knowledge that is required for the analogical argument to go through [4]. So, we can pin down Plato's *Meno* as the origins of a formal account of analogical reasoning.

Teach me half the ampliativeness
That thy powers must inferentially grow
Such surprise and hypotheses generation madness
From my reasonings would flow
The world will abductively know then, as I am abductively knowing now[8]
Adapted from Percy Bysshe Shelley, 'To a Skylark' (1820)

References

1. Bluck, R.S. (ed.): Plato's *Meno*. Cambridge University Press, Cambridge (1961)
2. Boter, G.J.: Discussion note: Plato *Meno* 82c2–3. Phronesis **33**(2), 208–215 (1988)
3. Cornford, F.M.: Plato's Theory of Knowledge: The Theaetetus and the Sophist translated with commentary. Dover Publications, Mineola (2003) (original, Liberal Arts Press, N.Y.: 1957)
4. Faller, M.: Plato's geometric logic. In: Proceedings of the Society for Ancient Greek Philosophy. http://polar.alaskapacific.edu/mfaller/PDF/MenoNWFnc.pdf (2003)

[8] Again, the words from the last verse of Shelley's 'To a Skylark' have been changed. [26].

5. Fann, K.T.: Peirce's Theory of Abduction. Martinus Nijhoff, The Hague (1970)
6. Goddu, G.C.: Book Review of Dov M. Gabay and John Woods, the reach of abduction: insight and trial. Informal Logic. **25**(3), 289–294 (2006)
7. Gulley, N.: Plato's Theory of Knowledge. Greenwood Press, Westport (1986)
8. Hegarty, M.: Diagrams in the mind and in the world: relations between external and internal visualizations. In: Blackwell, A., Mariott, K., Shimojima, A. (eds.) Diagramatic Representation and Inference, Proceedings of the Third International Conference, Diagrams 2004, Cambridge, UK, March 2004, pp. 1–12. Springer, Berlin (2004)
9. Job, R., Mulatti, C.: Do computational models of reasoning need a bit of semantics? In: Magnani, L., Li, P. (eds.) Model-Based Reasoning in Science, Technology, and Medicine, Series Studies in Computational Intelligence, vol 64, pp. 511–525. Springer, Berlin (2007)
10. Landry, E.: Recollection and the mathematician's method in Plato's *Meno*. Philosophia Mathematica **20**, 143–169 (2012)
11. Lehrer, K., Paxson, T.: Knowledge: undefeated justified true belief. J. Philos. **66**(8), 225–237 (1969)
12. Magnani, L.: Abduction, Reason and Science: Process of Discovery and Explanation. Kluwer, New York (2000)
13. Magnani, L.: Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Cognitive Systems Monographs, vol 3. Springer, Berlin (2009)
14. Magnani, L.: Email correspondence (14 June, 2012)
15. Moriarti, S.E.: Abduction and a theory of visual interpretation. Commun. Theory **6**(2), 167–187 (1996)
16. Mullins, P.: Peirce's abduction and Polanyi's tacit knowing. J. Specul. Philos. **16**(3), 198–224 (2002)
17. Nguifo, E.M.: Abduction and induction in learning task: which needs the other?. In: Flach, P., Kakas, A. (eds.) Proceedings of the ECAI'96 Workshop on Abductive and Inductive Reasoning, Budapest, Hungary, pp. 47–49, Wiley, London (1996)
18. Nubiola, J.: Abduction or the logic of surprise. Semiotica **153**(1–4), 117–130 (2005)
19. Olsen, S.: Plato, Proclus and Peirce: abduction and the foundations of the logic of discovery. In: Harris, R.B. (ed.) Neoplatonism and Contemporary Thought Part One, pp. 85–102. State University of New York Press, Albany (2002)
20. Peirce, C.S.: The Collected Papers of Charles Sanders Peirce (Harthshorne, C., Weiss, P., Burks, A. (eds.)) Harvard University Press, Cambridge (1931–1958)
21. Plato: Meno (Translated by W.K.C. Guthrie). In: Hamilton, E., Cairns, H. (eds.) The Collected Dialogues of Plato, pp. 353–384. Bollingen Series, Princeton University Press (1989)
22. Plato: Theaetetus. (Translated by F.M. Cornford). In: Hamilton, E., Cairns, H. (eds.) The Collected Dialogues of Plato, pp. 845–919. Bollingen Series, Princeton University Press, Princeton (1989)
23. Polanyi, M.: Knowing and Being: Essays by Michael Polanyi. (Greene, M. (ed.)) University of Chicago Press, Chicago (1969)
24. Scott, D.: Recollection and Experience: Plato's Theory of Learning and its Successors. Press Syndicate of the University of Cambridge, Cambridge (1995)
25. Sharples, R.W.: Discussion note: more on Plato, *Meno* 82c2–3. Phronesis **34**(2), 220–226 (1989)
26. Shelley, P.B.: To a skylark. http://www.bartleby.com/101/608.html
27. Stefanson, D.: Man as hero–hero as citizen: models of heroic thought and action in homer, Plato and Rousseau. PhD Thesis, University of Adelaide (2004). digital.library.adelaide.edu.au
28. Tiercelin, C.: Abduction and the semiotics of perception. Semiotica **153**(1–4), 389–412 (2005)

Generation of Hypotheses by Ampliation of Data

Emiliano Ippoliti

Abstract The issue of the ampliation of knowledge, in particular the generation of hypotheses by ampliation of data, is not effectively treatable from a logical point of view as it is affected by the multiplicity (nay infinity) of hypotheses that can be generated by the same data. This infinity is unavoidable as a consequence of the underdetermination of data by hypotheses. The paper argues that the issue of the generation of hypotheses is, instead, treatable from a heuristic viewpoint. More specifically the paper argues that the crucial step in the generation of hypotheses is the ampliation of data, that is the integration of the data of a problem with something not contained in them. The process of ampliation is crucial in the formation of hypotheses as it narrows the infinity of possible hypotheses that explain the data. It is essentially based on ampliative inferences, in particular analogies. The paper shows that there are three main ways to ampliate data and examines and compares two case studies of generation of hypotheses. The first one is the Black-Scholes-Merton equation, namely the hypothesis that the price of an option over time is given by the partial differential equation (PDE):

$$\frac{\partial O}{\partial t} + \frac{1}{2}\sigma^2 X^2 \frac{\partial^2 O}{\partial X^2} + rX\frac{\partial O}{\partial X} - rO = 0.$$

The second one is the generation of the Feynman Path Integral, a hypothesis about behaviour of quantum particles (about trajectories of quantum particles), namely the hypothesis that paths followed by electrons are all possible (infinite) paths, not just the 'classical' ones, which can be described by the functional integral:

$$K[a,b] = \int_a^b e^{{}^i/_\hbar S[a,b]} Dx(t)$$

E. Ippoliti (✉)
University of Rome La Sapienza, Rome, Italy
e-mail: emiliano.ippoliti@uniroma1.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_14, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

The issue of the methods of scientific discovery has been reignited in the last few decades in mathematics ([1, 2]), logic ([3–5]), natural science (see e.g. [6–14]), and in the social domain (see e.g. [15, 16]). The procedures for the generation of hypotheses from data are the crucial issue of this topic, since it is not effectively treatable from a logical point of view as it is affected by the multiplicity (nay infinity) of the relation between data and hypotheses. This infinity is unavoidable as a consequence of the underdetermination of data by hypotheses. The paper argues that the issue of the generation of hypotheses is, instead, treatable from a heuristic viewpoint. As a matter of fact heuristics are crucial in the formation of hypotheses as they narrow the infinity of paths leading to hypotheses that explain the data. They are essentially based on ampliative inferences and, in particular, analogies. The generation of hypotheses requires going beyond the data or, better, requires an 'integration of data' (see [17]) and in particular the ampliation of data with 'something' (properties, relations, entities) not contained in them at the beginning of the inferential process.[1] As a consequence, the paper supports the idea that the inductivist conception—that is the idea that data are sufficient to produce the hypotheses—is untenable. In effect, the multiplicity of hypotheses by data seems at least to show that the inductivist account of ampliation of knowledge, "according to which the data are all scientists need to infer a hypothesis" ([11], p. 35), is untenable. As a matter of fact, the thesis of underdetermination shows that data are not sufficient to generate hypotheses: it is necessary to add something to them in a non-deductive way, to integrate them (see [17]), in order to generate a hypothesis. It is necessary to go beyond data, to ampliate them in order to produce novelties. But this fact raises another problem, namely that "there are no rules for generating conceptually novel hypotheses from the data, nor [...] does having a specific problem tell scientists which facts are relevant. It is only when they have a testable hypothesis (problem solution) already in hand that they know which facts are relevant, namely, those which test the theory" ([11], p. 35).

Even though there are no rules in a strict sense, we have paths and methods for discovery ([5]); paths leading from data to hypotheses. In this paper I examine the ampliation of data. Ampliating data means introducing properties, relations and entities that are not contained in the data at the beginning of the inferential process. This allows us to extend (amplify) the set of properties holding for the entities involved in the problem and to reshape their relations. Furthermore, even though this extension is obviously uncertain, provisional and hypothetical, it is the only way to really generate conceptually novel hypotheses from the data. The problem here is how to individuate these properties. This can be done by means of ampliative inferences, which by definition are able to introduce something into the data that is not included in them at the beginning of the formation of hypotheses.

[1] The properties, relations and entities contained in the data at the beginning of the inferential process are the ones obtained by their preliminary analysis.

The paper analyses the issue by examining and comparing two case studies. Even though they belong to two different domains (finance and mathematical physics) the procedures are similar and can be compared; for this reason I believe that a lot can be learnt from them. The first one is the generation of the Black-Scholes-Merton equation, namely the hypothesis that the price of an option over time is given by the partial differential equation (PDE)

$$\frac{\partial O}{\partial t}+\frac{1}{2}\sigma^2X^2\frac{\partial^2 O}{\partial X^2}+rX\frac{\partial O}{\partial X}-rO=0$$

The second one is the generation of the Feynman Path Integral, a hypothesis about behaviour of quantum particles (about 'trajectories' of quantum particles), namely the hypothesis that paths followed by electrons are all possible (infinite) paths, not just the 'classical' ones. These paths can be described by the functional integral

$$K[a,b]=\int_a^b e^{i/_{\hbar}S[a,b]}Dx(t).$$

2 The Black-Scholes-Merton Hypothesis

The hypothesis put forward by the Black-Scholes-Merton equation is that the price of an option over time is given by the partial differential equation

$$\frac{\partial O}{\partial t}+\frac{1}{2}\sigma^2X^2\frac{\partial^2 O}{\partial X^2}+rX\frac{\partial O}{\partial X}-rO=0(BSM).$$

It is a very influential hypothesis about the valuation of stock options and was formulated in 1973 by Robert Merton (see [18]) on the basis of the works of Scholes and Black (see [19]). Let O denote an option (a financial security). Let r be the risk-free interest rate. Let X be the price of O, dependent on the price of the underlying stock at time t. Then we can write $O(X,t)$ to denote the price of an option. The BSM hypothesis states that, under the assumption of 'no arbitrage',[2] the option satisfies the above equation. This work was so successful and influential that it was awarded the Noble Price in economics in 1997, has generated new financial tools and has been widely applied to guarantees, insurance contracts and more generally to investment decisions in the last three decades. Moreover it still plays a pivotal role in financial economics.

The problem that the BSM equation aims at handling is at the core of financial economics: to determine when and for how much to buy or sell an option. This requires an appropriate level of risk in the financial transactions to be set and, in

[2] This means that it is not possible to take advantage of a price difference between two or more markets.

turn, it requires the price of an option at future given time *t* (the date of its maturity) to be determined. This forecast is obviously a crucial point, since it allow agents—who assume the risk and anticipate future revenues or payments—to ensure a profit above a certain level or insure themselves against a loss above a certain level. Hence, this price represents the value of an option. In order to make such an anticipation, it is necessary an interpretation of the available data—in this case the data are the prices and their changes; the historical series of prices of options. In particular it is necessary to ampliate them, ascribing properties that are not strictly contained in them.

In essence, the formula provides an answer to the issue of the efficient management of risk in financial economics: it gives a new method for determining the price of stock options that "virtually" eliminates risk, gradually reducing it by means of a calculation that is virtually the same made by the market itself. The solution offered by BSM equation, in short, is the creation of a risk-less portfolio by means of a variation of the basic strategic principle 'not to put all your eggs in one basket'. Such a risk-less portfolio is generated so that gradual changes in the composition of the portfolio reflect the gradual changes in the markets. In effect, gradual adjustments and variations in the composition of the portfolio are able to set a rate of return equal to the rate of return on any other 'risk-less' (or 'safe') financial instrument like the treasury bill (T.B.).[3] In this way, continuously an equilibrium between the portfolio and the global economics is generated. In effect, the financial risk does not disappear with this portfolio, but it is 'included', so to speak, in the stock price as the portfolio does not have a greater risk than the global economics.

Therefore the strategy of the BSM equation aims at avoiding and removing the most difficult task in financial markets, that is to foresee the price of a stock when the option matures. In effect, this approach relies on the idea that it is not necessary to make such a forecast, since "all you need to know is what the traders themselves know, the terms of the option (the strike price and time to expiration) and how volatile the stock is" ([20], p. 73). At this point you have two possibilities:

(1) the prices change within a narrow range, meaning that the stock is *stable*. Then the probability that its price will rise enough to make the options useful is very low;
(2) the stock is not stable, hence it is 'risky' since its prices vary a lot up or down. Thus, the probability that on one of those peaks the options will pay off is high. Then the options is valuable.

Of course, as the time of maturity approaches, the stock price changes and hence the relation between the option price and the share price is also modified. Therefore, in order to maintain a risk-free option-stock portfolio, the investors

[3] Since 1926 T.B. have paid an average of 3.8 % with very low risk both in good and bad times, while in the same interval of time the 500 stocks of S&P have paid an average of 13 %, but with great risks.

have to make gradual changes in its composition: (BSM) allows to calculate this changes and 'eliminate' the risk—just reducing it at the level of a *safe* financial tool.

The process of 'elimination' of risk can be illustrated by a simple example.[4] Let say that we have a derivative that gives us the right to buy one share in a certain firm at a strike price of $25 within the next four months. The value of this option depends on two parameters:

(i) the *strike price* (*sp*), that is the price at which the stock can be purchased (call option) or sold (put option). This is fixed at the time of formation of the option contract;
(ii) *today's stock price* (*tp*).

The difference between these two variables gives the profitability (or the value) of the option. The BSM equation simply conjectures that the higher is the stock price today, the greater is the chance that it will exceed $25 in the next four months, and, hence, the exercise of the option will be rewarding. For simplicity, let us assume that their ratio is 2, e.i. $sp/tp = 2$—e.g. if the stock price goes up by $4 today, then the option goes up by $2. An investor (owning a certain amount of shares in the firm) who wants to lower the risk of changes in the stock price can use a simple strategy that actually eliminates that risk: he can sell two options for every share that he owns (following the ratio of 2). At this point, according to this strategy, there are two possible scenarios about the risk of the portfolio:

(1) the portfolio created is risk-free: the invested capital must pay exactly the same return as the risk-free market interest rate on a four-month treasury bill-a typical 'risk-less' financial tool;
(2) If this it is not risk-free, arbitrage trading would begin to eliminate the possibility of making a risk-free profit.

As this example shows, the data (historical series of prices) as usual are not sufficient to produce a unique hypothesis about price behaviour, therefore they have to be ampliated and integrated with several properties and relations. In effect, the crucial step in the generation of the BSM equation is the ampliation of the data—that is the introduction of new properties to the entities of the data, in this case the introduction of new properties to prices. Several assumptions (on volatility, price changes, etc.) are needed to make the BSM model work.[5] I will discuss

[4] See The Prize in Economics 1997—Press Release. Nobelprize.org. 29 Aug 2012 http://www.nobelprize.org/nobel_prizes/economics/laureates/1997/press.html.

[5] The assumptions explicitly used in the BSM equation are:

(a) There is no arbitrage opportunity (i.e., there is no way to make a riskless profit).
(b) It is possible to borrow and lend cash at a known constant risk-free interest rate.
(c) It is possible to buy and sell any amount, even fractional, of stock (this includes short selling).
(d) The above transactions do not incur any fees or costs (i.e., frictionless market).
(e) The stock price follows a geometric Brownian motion, with constant drift and volatility.
(f) The underlying security does not pay a dividend.

the mode of generation of one of them that is crucial, namely the one stating that the price of the underlying asset of an option follows a geometric Brownian motion (GBM)[6] with constant drift and volatility. Expressed as a formula it holds that

$$\frac{\partial S}{S} = \mu dt + \sigma dW(GBM)$$

where W represents a Wiener process with increments dW having mean zero and variance dt.

In short, Brownian motion is a process that moves up and down in a random way, whose expected change over given time interval is equal to 0. It is like tossing a coin (a *random walk*), ending in the long run with more or less the same number of heads and tails. In particular GBM implies that the BSM equation ampliates and integrates prices (the data) with three properties:

(P1) *independence*, that is the idea that each price chance does not depend on the previous ones. Therefore past price changes do not affect those that follow (just as previous coin toss does not affect the outcome of the ones that follow). Accordingly, that means that "any information that could be used to predict tomorrow's price is contained in today's price, so there is no need to study the historical charts" ([20], p. 83).

(P2) *Stationarity*, that is the idea that the process moving the price changes remains the same over time: "if you assume coin tosses decide prices, the coin does not get switched or weighted in the middle of the game. All that changes are the number of heads or tails as the coin is tossed; not the coin itself" (Ibid.)

(P3) *Normal distribution*, that is the idea the price are distributed according to the bell curve, so that "most changes are small, an extremely few are large, in predictable and rapidly declining frequency" (Ibid.). Therefore the prices vary more or less uniformly over time and the size of most price-changes varies within a limited range, corresponding to the central portion of the bell curve. Moreover, the use of bell curve implies that prices are supposed to be distributed in a symmetric way: every 'up' in the prices is balanced by a 'down' in the long run, and these changes do not cluster together.

The forecasts about future price changes, and their frequency and magnitude, strictly depend on this three properties. For instance by (P1)–(P3) the range between the highest and lower prices should increase as the square root of the length of the elapsed time. This gives us the 'distance' that can be travelled by prices, how far the future prices can go from today's price (in probabilistic terms), so that we can bet on them.

[6] Brownian motion is a term borrowed from physics, where it describes the motion of a molecule in a uniformly warm medium. Bachelier (see [21]) was the first to conjecture that this process can describe price changes.

Moreover, the properties (P1)–(P3) tell us how the prices are related to each other. They specify (1) the modes that lead from a price to the next one and (2) what we should expect about future prices changes. They are establishing new connections and relations between the data that are not included in them at the beginning of the inferential process.

GBM is so important because it is a major factor in expressing as a formula the idea that the market will eliminate the risk (by making the same calculation of the BSM equation). Under this respect, it is a straight consequence of the Efficient Market Hypothesis (EFH). EFH, in short, states that the prices fully reflect all relevant information, that the random walk is the best approximation to describe such markets and that you cannot beat such an unpredictable market. This is a strict neoclassical assumption and therefore BSM is a way to explain the data (and make forecasts) compatible with neoclassical theory. In other words, the equation is nothing else that a way to rewrite mathematically a set of neoclassical assumptions.

Nevertheless the point of my paper is not to question the nature of the hypotheses behind the BSM equation (it has been done enough, see e.g. [20]), but to examine and question the procedure of ampliation and integration of the data employed in the generation of BSM. These procedures are responsible for selecting and supporting certain properties, like GBM, as compatible and cogent with the data.

In the case of GBM, the integration of the prices (the data) is based on an ampliative inference, namely an analogy. As a matter of fact, we have that:

(i) the returns predicted by GBM are independent of the value of the process (stock price) - and in effect that is what is expected according to the neoclassical theory;
(ii) just like real stock prices GBM has only positive values;
(iii) GBM displays the same ‘roughness’ in its paths as seen in the paths of stock prices;
(iv) GBM is among the simplest of the continuous-time probabilistic processes.

These similarities are the basis for the selection of GBM as a candidate for modelling the prices changes in financial markets (see [21, 22]). The crucial property about prices is integrated in the data by means of an analogy over (i)-(iv).

3 The Feynman Path Integrals Hypothesis

The second example is the generation of the Feynman Path Integral, a hypothesis about the behaviour of quantum particles—in particular about the paths, or in a sense the ‘trajectories’, of quantum particles. It aims at solving the problems of QM data by starting from the problem of the infinite self-energies in the classical theory. The hypothesis simply states that in order to determine the probability to detect an electron at a given location a and time t_a you have to consider all the possible (infinite) paths followed by electrons, not just the ‘classical’ ones. In order to implement this hypothesis it is necessary to carry out integrals over all space variables at every instant of time, generating the well-known functional integral:

$$K[a,b] = \int_a^b e^{i/_{\hbar} S[a,b]} Dx(t)(FPI).$$

FPI allows to calculate the probability $P(a,b)$ that an electron located at point a and time t_a the location b and time t_b. In particular $P(a,b) = |K(a,b)|^2$. Here $K(a,b)$ is the amplitude to go from a to b. This amplitude is the sum of contributions of the amplitude of each path—which has a phase proportional to the action S. The total amplitude is given by the FPI. The hypothesis was put forward by Richard Feynman in 1948 (see [23]) and since its first formulation was so ill-defined and controversial that Feynman himself doubted its mathematical legitimacy. As a matter of fact he noted that:

> The necessity to redefine the method of integration does not destroy the concept of integration. So we feel that the possible awkwardness of the special definition of the sum over all paths. . .may eventually require new definitions to be formulated. Nevertheless, the concept of the sum over all paths, like the concept of an ordinary integral, is independent of the special definition and valid in spite of the failure of such definitions ([24], p. 6).

As well known a fundamental difference between classical physics and quantum theory is the fact that in the latter the predictions about certain variables, like trajectories, can only be made in probabilistic terms. In classical physics, instead, we can give a definite and precise answer: to determining a trajectory of an entity we need the its initial velocity and the forces acting on it.

Since in QM we can only calculate the probability that a particle starting at a given location i and time t_i will reach location d at time t_d, the classical notion of trajectory has been removed (in the traditional formulation). In order to make sense of the data of QM dynamics (like the double split experiment) a lot of hypotheses can be put forward (see e.g. [24] p. 6). As usual, the data alone cannot resolve the dispute. You can integrate them in a lot of different ways, they admit infinite paths leading to the hypotheses.

The standard treatment of QM data—in this case an electron passing through two slits and creating an interference pattern on a screen—employs (a) the hypothesis of the wave nature of the electron and (b) Huygens’ principle to calculate the interference of elementary waves which come from the two slits. Feynman offered an alternative approach to describe the behaviour of QM entities and the path integral (or *sum over histories*) is the tool to compute it. Feynman conjectured that an electron can be thought as a particle able to cross both paths that go through the slits. So he approached the issue in classical terms.

In particular the solution offered by FPI is such that you can in fact use again the concept of trajectory in order to make sense of the QM data: simply you have to consider all the possible trajectories, connecting two locations a and b, both in terms of space and time (i.e. velocity). The issue here is that they are infinite in number: there are infinite many paths connecting two given points and each of these paths can be travelled with infinite many velocities. These infinite possibilities are also known as the 'histories' of the system. In order to compute the probability to detect an electron at the final location, roughly, a 'number' is associated with each of these infinite possibilities and the integral is calculated over them—that is the 'numbers' are added. This number is the so called 'amplitude'-a- and the sum of the amplitude of each path gives the total amplitude. Each trajectory is associated with a norm unit amplitude so that each path gives the same contribution, in an egalitarian fashion. Then the square of the mod of the total amplitude, $|a|^2$, gives the probability to detect the particle at a given location. In effect during the integration (sum) some numbers cancel each other while others are added, so that the result of this 'sum' gives the probability to detect the particle at the given location. In other words, the path integral approach 'splits' the temporal evolution of the system into tiny intervals so that the overall evolution is given by the product of these small intervals. More specifically, the techniques employed are *time-slicing* and *space-slicing*: the space is sliced into smaller and smaller interval over all the possible trajectories and the time is sliced considering all the possible velocities along the paths.

The path integral is obtained by taking the limit of the time intervals going to zero, and the limit of steps (space evolution) going to infinity. In this way we have a new type of integration, i.e. an integral over all space variables at every instant of time. Here the approximation to the overall evolution of the system is obtained by zig-zag paths (and thus in a linear way), even though obviously this is not the only way to approximate an integral.

Nevertheless, as I have underlined, this procedure is mathematically controversial and not rigorous. From a formal point of view, FPI is an integral over an infinite dimensional function space: not only to define a reasonable measure on such a space is problematic, but also the general properties of this kind of space are hard to understand. In effect, when a sum over an infinite number of objects is required, typically it is possible to introduce a measure to put on the space of these objects. But in this case, if you admit that the for path-space a notion of dimension exists, it must be infinite-dimensional. Accordingly, it's hard to see how it is possible to define—for any plausible notion of distance—a measure over such a space that could be mathematically treatable. Just to give the idea of the puzzle, every sphere will fail to be compact in such infinite-dimensional space (!).

A possible way to solve this puzzle is to use the Lebesgue's measure, but "the Lebesgue-type flat measure D γ on a space of paths is not defined from a mathematical point of view and cannot be used as a reference measure" ([25]). Moreover, Cameron ([26]) "showed that no such countably additive measure exists. To date, there is no measure-theoretic definition of the Feynman integral. There are, however, several definitions of it which do not use measure" ([27], p. 457). Therefore, "since it is not possible to give meaning to the Feynman integral $I(f)$ of a function f on the space Γ of paths γ in terms of an integral with respect to a σ—additive (complex-valued) measure, one can try to define $I(f)$ as a linear continuous functional on a suitable linear space of functions f" ([25]). But, again, this does not solve definitively the problem and the calculation of this integral can be very hard. The same holds for other approaches aiming at offering a mathematical foundation of FPI.

Just to give an example, let us consider the calculation by means of FPI of the probability that certain particles will interact with each other in a given way. In this case a mathematical 'trick' is required: every time that in Feynman's formula the time coordinate t occurs, it is necessary to introduce the extra factor i - the imaginary unit. After the path integral is computed, you have to invert this substitution. "The replacement might seem artificial and implausible. In a way, it corresponds to transforming the time coordinate into just another space coordinate. Fact is, it makes the Feynman recipe give the right answers. There's even an exact proof, found by two mathematical physicists, Konrad Osterwalder from Switzerland and the German Robert Schrader: They proved a theorem showing that the properties of a quantum theory formulated in the space-time of special relativity can indeed be reconstructed exactly by using the Feynman recipe on an imaginary-time version of that same space-time" ([28]).

As it is often underlined (see e.g. [25]), Feynman introduced the path integral in a heuristic way—that is in virtue of the right calculations and predictions it offers—and he made sense of this mathematical entity by means of physical considerations when the mathematics and the calculation became unclear and uncertain.

It is essential to note that also the generation of the FPI hypothesis is based on a process of integration of the data guided by an ampliative inference. In particular the process of its generation was guided by two constraints: the avoidance of infinite self-energies and the search for a link between quantum and classical systems. Under these restrictions, an analogy is at core of the hypothesis of path integral. The analogy is based on the available knowledge about classical systems dynamics—in particular the Least Action Principle (LAP), which states that the classical trajectory connecting two given points in the space is the one minimizing the action S. Following a suggestion of Dirac, in effect Feynman tried to import a 'version' of LPA onto the QM entities, in order to make sense of their behaviour in terms of classical physical notions and to shed light on how "classical mechanics could naturally arise as a special case of quantum mechanics when h was allowed to go to zero" ([24], p. vii–viii).

Therefore the crucial step in the generation of FPI hypothesis is the introduction of new properties to data. In this case, remarkably, we have that the classical notion of trajectory is re-introduced into the electrons by means of LPA. In this way new entities and relations are introduced in the data. So "Feynman's approach is particularly suggestive as it creates a bridge between the classical Lagrangian description of the physical world and the quantum one, reintroducing in quantum mechanics the classical concept of trajectory, which had been banned by the traditional formulation of the theory. It allows, at least heuristically, to associate a quantum evolution to each classical Lagrangian" ([25]).

As a consequence of the successful predictions and calculations of path integrals, Feynman formulated the following assumptions (or axioms):

F1. the probability for an event is given by the squared length of a complex number called the probability amplitude;

F2. the probability amplitude is given by adding together the contributions of all the histories in configuration space;

F3. the contribution of a history to the amplitude is proportional to $e^{iS/\hbar}$, where S is the action of that history;

4 Processes of Ampliation of Data

The central issue in the generation of hypotheses, as shown by the two examples, is how to select and support the properties that are ascribed to the entities of the problem, how to model the process of integration of data. It is possible to identify three main cases of ampliation of data, that is kinds of properties that can be ascribed to the concepts and entities described by the data:

(1) the properties are already known and well established, as they hold for other known entities. For instance the property of continuity was transferred from mathematics to Newtonian Mechanics in order to treat the concept of physical time. In this case the properties are ascribed to the objects described by the data by means of hybridizations and interpretations. This operation is not mechanical and often requires that new definitions, entities and relations are created in the domain of the problem. They are, so to speak, 'solutions in search of problems': the strategy is based on the idea to use known results to solve existing problems. The GBM in the BSM equation is, in a precise sense, an example of this way of ampliation.

(2) The properties eligible for the ampliation of the data are highly hypothetical and controversial, but already formulated and employed in some domain. A remarkable example (see [29–31]) is the crucial hypothesis employed by Hippocrates of Chios to solve the problem of the quadrature of the lune, namely the hypothesis that the areas of two circles or semicircles are to each other as the squares on their diameters. The property expressed by this proposition (which became the prop XII.2 in Euclid's *Elements*) has already been formulated and known but it was unproven, "for there is widespread doubt that Hippocrates actually had a valid proof" ([29], p. 18).

(3) The properties eligible for the ampliation of the data are formulated for the first time during the generation of the hypothesis. The hypothesis generated solves a specific problem and represents a genuine theoretical novelty. Notable examples are the Feynman Path Integral and Leonard Euler's solution of Mengoli's problem (see [31, 32] Chap. 5).

Obviously each ampliation of data is unsure: it requires a 'leap in the dark' and it is never safe beyond any doubt. Nevertheless we can distinguish between good and bad leaps by looking at the procedures employed to ampliate the data that lead to the generation of the hypotheses.

In the case of the BSM equation I argue that the ampliation of data is put forward using a set of ‘bad’ reasons and that, accordingly, the whole process of integration of data is misleading and ill-structured. It follows that the properties ascribed to the data by ampliative inference are generated and selected erroneously.

This sheds light on the reasons for the implausibility of the hypotheses of the BSM equation—*why* these hypotheses are untenable, not simply that they are unrealistic or ‘shaky’. In essence, I question the processes of ampliation of data that generate the hypotheses and, hence, their plausibility.

That fact the BSM equation is inadequate to manage the problem of risk management because the hypotheses it adopts are completely implausible is a well established and clear fact. The same use of this formula in real-world finance has been questioned (see [33]). For instance, Mandelbrot says that

> the whole edifice hung together—provided you assume Bachelier and his latter-day disciples are correct. Variance and standard deviation are good proxies for risk, as Markowitz posited—provided the bell curve correctly describes how prices move. Sharpe’s beta and cost-of-capital estimates make sense - provided Markowitz is right and, in turn, Bachelier is right. And Black-Scholes is right—again, provided you assume the bell curve is relevant and that prices move continuously. Taken together, this intellectual edifice is an extraordinary testament to human ingenuity ([20], p. 76–77).

Less clear, instead, is the process of generation of the hypotheses and why it is faulty. Well, first of all, it is worth noting that the similarities employed to propose GMB are really meagre and weak, and the consequent theoretical construction as a whole “is no stronger than its weakest member” (Mandelbrot 2004, p. 77). Moreover, the integration of data is conducted in a top-down fashion: the hypothesis is the neoclassical one, and thus given, and the mathematical model follows from it. Furthermore, the process of generation of hypothesis ‘starts with algebra’ (the mathematical concepts and tools - just like GBM - are already known and given) and it is based on the search of ‘rigor’—the compatibility or, better, the derivability of the known mathematical model from the economic orthodoxy, i.e. the neoclassical framework and its focus on the determination of points of equilibrium).

Nevertheless this strategy is flawed and then the BSM equation is only seemingly sound and effective, while in fact it uses a trick—the accommodation of data with the neoclassical hypotheses. In fact the data are ‘sacrificed’ to hypotheses, in the sense that a part of the data (the extreme events and their frequency) is deliberately ignored in order to keep the rest of the data compatible with the neoclassical framework. The aim of the inferential process is to ‘keep’ and ‘save’ the hypotheses of mainstream economics. Thus, the process of generation of hypothesis is carried out in such a way that you are seeking for the best available mathematics that approximates the hypotheses, not the best hypothesis for the all the data. In essence, the BSM equation is generated in such a way that the problem and the data are interpreted as to conveniently fit the neoclassical hypothesis.

Hence, not simply here we have a solution in search of a problem, but the problem and the data are modified in order to be treatable by some known and convenient '*solution*'.

No surprise then that the crucial problem in the integration of the GBM property to prices is that it leads us to overlook and ignore a part of the data—namely the chance of large price changes (the so called *fat tails*). In effect a lot of historical prices series show too many very big price changes, too many very small changes and not enough medium-sized ones. Moreover the series show a lot of clusters (concentration is a small interval of time of big changes), which are not properly the signs of a 'random walk'. Thus, convenience in the mathematics and not a cogent integration of data is the reason leading to certain hypotheses like GBM in the BSM equation.

Moreover, this critical mistake is such that BSM ignores or cut off significant data and connections between the data and, in the end, generates wrong predictions. The market crashes in 1897 and 2008 are just two notable examples in this sense: a bad assessment of financial risk has led the markets to a bad management of risk, which in turn generated the crashes. And the BSM equation has been the standard tool for the evaluation of risk in the last four decades. So, a seemingly rigorous and sound hypothesis turned out to be completely unsound and misleading. In essence, it does not solve any problem at all and on the contrary it creates problems in real life.

In order to properly and better explain the data, in effect, you need different and new mathematical concepts and tools. And it is just what, for example, Benoit Mandelbrot did. He elaborated new concepts - *scaling*, *dependence*, *fractality* - and created a new mathematics[7] in order to make sense of the data, generating a different 'vision' of their connections and relations. This vision is not definitive, but it is reliable and plausible (compatible with the data) and it is able to produce a genuine advancement in our knowledge and understanding of financial markets.

As concerns the generation of PFI, I have shown that the integration of data in this case is bottom-up (that is from data up to the hypotheses), based on heuristics and fertility (and not rigor), and 'ends up' with algebra (that is a new mathematical entity is produced). Furthermore, since it is generated by analogy with classical physics (the Least Action Principle), FPI requires the production of new assumptions (the Feynman postulates (F1)-(F3)) in order to give a cogent interpretation of the data. The method used by Feynman to generate FPI is very similar to the one employed by Leibniz and Newton in the generation of calculus and the relation between data and hypotheses is reversed whit respect to the BSM equation. The hypothesis generated is so controversial and bold just because it does not cut off data and aim at integrating them all, by creating new concepts and mathematical entities. Tellingly, the axioms are formulated because they allow to reach the wanted result (the calculation of the position of an electron that avoids infinite

[7] Fractal mathematics originates just from a work on cotton pricing.

self-energies) even though they may seem absurd or counter-intuitive (as the sum over all paths can appear at first sight).

In this case, hypotheses can be—and in effect are—sacrificed to data, in the sense that hypotheses can be abandoned or changed if they are not compatible with even a small part of the data. In the end, the example shows that the existence and rigor of certain mathematical entities is a minor issue compared to their heuristic potentiality.

Even though controversial, non rigorous and bold, FPI has revealed fruitful and fertile as it has achieved many successes and has a lot of advantages in complex problems. I mention here just few example. First of all, FPI have a compact formal expressions that involve integrals over numbers rather than operators, and this allows the application of familiar approximation techniques (e.g. the method of steepest descents or stationary phase) to issues like mean field approximation (in case of small fluctuations). Furthermore PFI show how a quantum field theory in S+1 space-time dimensions (S spaces and 1 time) is connected with the statistical mechanics of a system in S+1 spatial dimensions by means of the analytical continuation of the time dimension (the ‘Wick rotations’). This link has originated the treatment of the field theories by means of statistical mechanics and the renormalization group introduced by Wilson (*lattice theories*). Again, in string theory, the probability of interactions of certain strings can be calculated as a path integral by summing up not only all the possibilities that a string can travel through space, but also all the possible deformations that a string can have on the way.

5 Conclusion

As I have shown, the processes of generation of the two hypotheses BSM and FPI is very different, and so are their stories and outcomes.

On one side we have that the BSM equation is generated in a top-down fashion (that is starting from the hypotheses and going down to the data) and in a *seemingly* rigorous way by employing a convenient and well-established mathematics. During the process of generation, the hypotheses are the fixed points and data can be ignored or cut off if they don’t fit the theoretical framework. Moreover, it seems to be perfectly rigorous and sound, so that at the beginning it was received as an absolute breakthrough, and awarded the Nobel prize, but ended in a complete overthrow. Introduced as the ultimate and *general* tool to manage financial risk, in fact it don’t solve the problem but is simply a way to rewrite in a mathematical style the neoclassical hypotheses about markets. In this sense it provided an integration of the data that don’t ampliate at all the neoclassical framework, but it is simply a way to make them compatible with the mainstream economics. There is nothing outside the mainstream economics coming into the data.

On the other side, FPI is generated in a bottom-up fashion, supported by its fruitfulness and fertility rather than its (lack of) rigor. Its main aim is the cogent integration of the data by means of Least Action Principle, so that the data are the fixed points and the hypotheses can be modified or cut off at any moment if they don't fit the entire set of data. It holds obviously for (F1)-(F3), which have been formulated in that specific form as they are good at making sense of the data. Thus, PFI really solves the problem and was developed as a *local* tool to do it. Moreover it produced new mathematical entities and concepts.

For a period of time the hypothesis FPI was regarded suspiciously, considered at last as an interesting reformulation of QM, but in the end Feynman was awarded the Nobel Prize and today, even though still controversial and *ill-founded*, it helps to solve interesting and complex problems in several domains (including finance, see [34]). Tellingly, from a formal point of view, BSM equation is nothing else but a special case of a PDE that can be solved by using the Feynman-Kac formula, a variant of Feynman Path Integral.

References

1. Cellucci, C.: Filosofia e matematica. Laterza, Roma (2002)
2. Polya, G.: Mathematics and Plausible Reasoning. Princeton University Press, Princeton (1954)
3. Magnani, L.: Abduction, Reason, and Science. Processes of Discovery and Explanation. Kluwer Academic, New York (2001)
4. Magnani, L., Carnielli, W., Pizzi, C. (eds.): Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery. Springer, Heidelberg (2010)
5. Cellucci, C.: Perch© ancora la filosofia. Laterza, Roma (2008)
6. Hanson, N.: Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science. Cambridge University Press, Cambridge (1958)
7. Laudan, L.: Progress and its Problems. University of California Press, Berkeley and LA (1977)
8. Simon, H.: Models of Discovery. Reidel, Dordrecht (1977)
9. Simon, H., Langley, P., Bradshaw, G., Zytkow, J. (eds.): Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press, Boston (1987)
10. Nickles, T. (ed.): Scientific discovery: Logic and Rationality. Springer, Boston (1980)
11. Nickles, T. (ed.): Scientific discovery: Case Studies. Springer, Boston (1980)
12. Nickles, T., Meheus, J. (eds.): Methods of Discovery and Creativity. Springer, New York (2009)
13. Grosholz, E., Breger, H. (eds.): The Growth of Mathematical Knowledge. Springer, Dordercht (2000)
14. Darden, L. (ed.): Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution. Cambridge University Press, new York (2006)
15. Abbott, A.: Method of Discovery. W.W. Norton & Company Inc., New York (2004)
16. Hendry, D.F.: Empirical economic model discovery and theory evaluation. Ration. Market Moral **2**, 115–145 (2011)
17. Ippoliti, E.: Between data and hypotheses. In: Cellucci, C., Grosholz, E., Ippoliti, E. (eds.) Logic and knowledge, pp. 237–262. Cambridge Scholars Publishing, Newcastle Upon Tyne (2011)
18. Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**, 141–183 (1973)
19. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**, 637–654 (1973)
20. Mandelbrot, B.: The (mis)Behavior of Markets. Basic Books, New York (2004)

21. Bachelier, L.: Théorie de la spéculation. Annales Scientifiques de l'cole Normale Supérieure **17**(1), 21–86 (1900)
22. Black, F.: How we came up with the option formula. J. Portfolio Manag. **15**, 4–8 (1989)
23. Feynman, R.: Princeton University Press, Princeton (1985)
24. Feynman, R., Hibbs, A.: Quantum Mechanics and Path Integral. McGraw-Hill, New York (1965)
25. Albeverio, S., Mazzucchi, S.: Path integral: mathematical aspects. Scholarpedia **6**(1), 8832 (2011)
26. Cameron, R.H.: A family of integrals serving to connect the wiener and feynman integrals. J. Math. Phys. **39**, 126–140 (1960)
27. Keller, J., McLaughlin, D.: The feynman integral. Am. Math. Mon. **82**, 451–465 (1975)
28. Poessel, M.: The sum over all possibilities: the path integral formulation of quantum theory. Einstein, Online **2** (2010)
29. Dunham, W.: Journey Through Genius: The Great Theorems of Mathematics. Wiley, New York (1990)
30. Cellucci, C.: Le ragione della logica. Laterza, Bari (1998)
31. Ippoliti, E.: Demonstrative and non-demonstrative reasoning by analogy. In: Cellucci, C., Pecere, P. (eds.): Demonstrative and Non-Demonstrative Reasoning by Analogy in Mathematics and Natural Science, pp. 307–338. Cassino University Press, cassino (2006)
32. Ippoliti, E.: Il vero e il plausibile. Lulu, Morrisville(N.C.) (2007)
33. Haug, E.G., Taleb, N.N.: Option traders use (very) sophisticated heuristics, never the black-scholes-merton formula. J. Econ. Behav. Organ. **77**(2), 97–106 (2011)
34. Kleinert, H.: Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets. World Scientific, Singapore (2004)
35. Feynman, R.: Space-time approach to non-relativistic quantum mechanics. Rev. Mod. Phys. **20**(2), 367–387 (1948)

Abduction, Competing Models and the Virtues of Hypotheses

H. G. Callaway

Abstract This paper focuses on abduction as explicit or readily formulatable inference to possible explanatory hypotheses—as contrasted with inference to conceptual innovations or abductive logic as a cycle of hypothesis, deduction of consequences and inductive testing. Inference to an explanation is often a matter of projection or extrapolation of elements of accepted theory for the solution of outstanding problems in particular domains of inquiry. I say, "projection or extrapolation" of accepted theory, but I mean to point to something broader and suggest how elements of accepted theory constrain emergent models and plausible inferences to explanations—in a quasi-rationalistic fashion. I draw illustrations from quantum gravity below just because there is so little direct evidence available in the field. It is in such cases that Peirce's discussions of abduction provide the most plausible support for the idea of a logic of abduction—as inference to readily formulatable explanatory hypotheses. The possible need for conceptual innovation points to limits on the possibility of a logic of abduction of a *more* rationalistic character—selecting uniquely superior explanations. Abduction conceived as a repeating cycle of inquiry also points to limits on our expectations for an abductive logic. My chief point is that the character of inference to an explanation, viewed below as embedded within arguments from analogy, is so little compelling, as a matter of logical form alone, that there will always be a pluralism of plausible alternatives among untested hypotheses and inferences to them—calling for some comparative evaluation. This point will take us to some consideration of the virtues of hypotheses—as a description of the range of this pluralism.

H. G. Callaway (✉)
Philadelphia, PA, USA
e-mail: HGCallaway@live.com

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_15, © Springer-Verlag Berlin Heidelberg 2014

1 Conceptual Constraint on Hypothesis

The constraint of accepted theory and pre-existing conceptual resources upon inference to explanatory hypothesis cannot be absolute. In proposing a new explanatory hypothesis, one typically seeks to meet problems in accepted theory, account for some anomaly, by way of supplementing and/or modifying theory or elements of belief, and these proposals may involve innovations in the conceptual resources of the theory. Yet, this is far from saying that the pre-existing theory in a domain and the conceptual resources employed in it will have no influence on plausible inferences to explanatory hypotheses.

Consider the abstract, bare-bones scheme which Peirce provides for abductive inference in his late writings[1]:

> 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.

I will pass for the moment on the question of whether such inference requires a distinctive logic, properly so called, for its elucidation, though I acknowledge that this kind of inference welcomes elucidation and that similar inferences form a class concerning which study is advised. Whether or not there is a logic of abduction,[2] I think there is an *art* of inference to explanatory hypotheses and that this is worthy of study and attention. Still the plausibility of acknowledging an art of inference to hypotheses is part of the plausibility of denying a distinctive logic of abduction. As formulated here, abduction is a matter of cognitive expectations. First of all, some fact C is said to be "surprising," which is to say that C in some manner fails to accord with established expectations. In the second premise, the supposition is that if the hypothesis A were true, then "C would be a matter of course," which is to say that C would no longer be surprising, but instead would accord with expectations arising in consequence of the supposed truth of hypothesis A. What lends support to hypothesis A in the conclusion is that the surprising character of C would be removed and C would become consistent with envisaged expectations arising on the supposition of the hypothesis. To evaluate this form of inference we have to know whether and how the truth of the premises would render the conclusion plausible. What, then, changes the relevant expectations?

The first requirement is to understand what it is that creates the expectations, including the initially surprising character of fact C and the removal of the surprising

[1] See Peirce, C.S. (1903) [17] "Pragmatism as the Logic of Abduction," in *The Essential Peirce, Selected Philosophical Writings*, Vol. 2 (1893–1913), p. 231.

[2] Regarding alternative approaches to the topic of the logic of abduction, see Aliseda, A. (2006) [1] *Abductive Reasoning. Logical Investigations into Discovery and Explanation*; Gabbay, D.M. and J. Woods (2005) [7] *The Reach of Abduction, A Practical Logic of Cognitive Systems*, 2 vols.; and Magnani, L. (2009) [14] *Abductive Cognition, The Eco-Cognitive Dimensions of Hypothetical Reasoning.*

character on the supposition of hypothesis A. The answer which seems most reasonable is that in the two cases, we envisage or presuppose some accepted, or prospective, context of theoretical understanding (or belief) relevant to the fact C, of which C is at first not, and afterward becomes, an expected part. We might imagine, for instance, that fact C presents a counter-example to an accepted theory T_1 over domain D, or that it is not encompassed by theory T_1; and in consequence, fact C is surprising to those whose expectations regarding domain D are structured by their acceptance of theory T_1. Correspondingly, to say that if hypothesis A were accepted, then "C would be a matter of course," suggests that there is some alternative formulatable theory T_2, including A, over domain D, such that in the simplest case, C or a related conditional with C as consequent, is a logical implication of theory T_2. In short, the idea is that it is always a theoretical context, or (weakly or firmly held) beliefs (which can be idealized as a theory), including some typical patterns of inference among its concepts or terms, which structures and is suited to create or remove the conceptually relevant expectations involved in the "surprising" or "matter of course" character of particular observed facts of experience.

This is to say that given particular intellectual configurations involving accepted belief or theory and problematic evidence of facts relevant to, or included within, the same domain as the accepted theory—such configurations—will always create some needed conceptual constraint upon possible hypotheses designed or functioning to deal with the problematic aspects of the theoretical situation. For, it is the concepts and patterns of inference of some particular theory, or idealization of existing belief, which renders some fact C "surprising" in the first place, and plausible constraint upon alternative hypotheses and corresponding alternative belief or theory must seek to preserve relevant patterns of expectation while removing the particular surprise. Preserving relevant patterns of expectation, so far as possible, means preserving the explanatory accomplishment of the theory or belief system theretofore accepted. Here we need to notice, too, that consistent with the argument for hypothesis A given above, we may reasonably suppose that any of several alternative arguments are open to us:

> 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.

So long as we deal with these matters in this abstract fashion, we can consistently suppose that we have any number of similar possible inferences, each putting forth some particular possible hypothesis, A', A'', A''', etc. The pattern or generalized form of inference Peirce ascribes to abduction is not plausibly regarded as a matter of inference to some unique "best explanation," and is instead better understood as a pattern of inference to any of several possible explanations. This in turn suggests that it is not the *pattern of inference* which "wears the pants," in actual and successful inferences of this sort, and that we have to access the actual content and concrete problems of particular fields of inquiry in order to make much sense out of inference to explanatory hypotheses. In short, we need to examine, and depend upon, the actual patterns of conceptual expectations arising in particular contexts of inquiry: it is these concrete expectations which "wear the pants" in any actual inference to an explanation.

It is reasonable to expect such a result, and perhaps the point will not be very widely questioned. Explanations are always embedded in particular conceptual frameworks or theoretical systems. If a virus V is the cause of disease D, say, the common cold, and we can, in fact, re-describe virus V, as the last thing mentioned by Jones, then it follows that the last thing mentioned by Jones is the cause of disease D. More generally, if A is the cause of B, then A remains the cause of B under any alternative description which picks A out. However, an explanation depends on particular conceptual resources and theoretical context, an explanation makes use of preferred or salient forms of description linked by theoretical context to particular consequences; and alternative characterizations of the things we refer to in explanations do not carry the same explanatory force or weight. We might explain someone catching cold by contact with virus V, but it won't do to offer explanation in terms of contact with the last thing mentioned by Jones. Our accepted theoretical concepts have a certain inherited salience or grounded projectability, due to their roles in successful explanation and their comprehension of a range of evidence. We naturally want to preserve them so far as possible in the problematic situation. Regarding things mentioned by Jones, in general, we have no relevant generalizations.

There will always be a range of explanatory innovations that may be proposed, regarding unsolved problems, running from the more conservative to the less conservative; and it is important, in ruling out "wild guessing," that attention be initially directed to more conservative proposals. To say that there is at least a quasi-logic, or art of abduction, is to resist over-emphasis on the idea that "hypothesis is guesswork" or that it is merely guesswork. If there is an important distinction between "guesswork" and "enlightened guess work," as Quine and Ullian maintain, in *The Web of Belief*, then this points to the possibility of systematic and comparative evaluation of the better and worse among unverified hypotheses[3]—thought alternative abductive arguments share the same abstract form.

2 Quantum Gravity, Analogy and Hypothesis

Contemporary inquiries and discussions of the relationship between general relativity and quantum mechanics are of particular interest in the present context, just because empirical evidence has been so scarce. The gravitational interaction, though cumulative over long distances, is much weaker than any of the other interactions, and because of this it is exceedingly difficult to institute experiments

[3] "In a word, hypothesis is guesswork," say Quine and Ullian (1978) [24] in *The Web of Belief*, second ed., p. 65, "but it can be enlightened guesswork."

sufficiently delicate to measure effects requiring the precision of a quantum theory of gravitation.[4] The conflicts between general relativity and quantum mechanics are of a theoretical rather than a more empirical nature: basic in this is the conflict between the smooth continuities of the gravitational field of relativistic space-time in contrast to the bubbling discreteness of quantum fields. As Brian Greene has put the point, "The notion of a smooth spatial geometry, the central principle of general relativity, is destroyed by the violent fluctuations of the quantum world on short distance scales."[5]

In *Objective Knowledge* (1979), Karl Popper provided a point of reference in opposition to methodological conservatism when he characterizes his conception of conjecture in relation to "the method of science": "*The method of science*," he says, *"is the method of bold conjectures and ingenious and severe attempts to refute them*;"[6] This implies that in the comparative evaluation of a pair of new hypotheses, as possible modifications of some pre-existing theory, we should generally prefer that new theory, containing an hypothesis, which has greater logical comprehension, implying a larger range of testable consequences. Einstein's bold innovations regarding space-time and matter in motion are a plausible model here.

While Stephen Hawking expresses sympathy for Popper's falsificationism, he says, too, that "In practice, it seems that one develops a new theory which in truth is only an extension of the old."[7] Brian Greene writes that "rather than trying through one leap, to incorporate all we know about the physical universe in developing a new theory, it is often far more profitable to take many small steps that sequentially include the newest discoveries from the forefront of research."[8] This claim compares more favorably with Popper's emphasis on "trial and error" and the "piecemeal approach," in *The Poverty of Historicism* (1957) and equally with Popper's emphasis on simplicity in *The Open Universe* (1982).[9] "The method of science", he says there, "depends upon our attempts to describe the world in simple theories." In contemporary physics, the beauty and generality of Einstein's physics of space-time and gravitation seems to be at war with the precision and predictive success of quantum mechanics; and Einstein's later dream of a unified field theory, though arguably more plausible during Einstein's years at Princeton—as a conservative modification of his physics of gravitation, space-time and matter aiming to integrate electromagnetism—now counts as a proposal far wide of the mark, since the quantum-mechanical approach has subsequently encompassed three of the four fundamental forces from the opposite direction—as a quantum field theory. What

[4] See, e.g., Feynman, Richard (1985) [6] *QED: The Strange Theory of Light and Matter*, p. 151.

[5] Greene, Brian (1999) [9] *The Elegant Universe*, p. 129.

[6] Popper, Karl (1979) [23] *Objective Knowledge*, revised edition, p. 81.

[7] See Hawking, Stephen (2006) [11] *A Briefer History of Time*, p. 20 in the German edition.

[8] Greene (1999) [9] *The Elegant Universe*, p. 121.

[9] See Popper, Karl (1957) [21] *The Poverty of Historicism*, p. 75; Popper, Karl (1982) [22] *The Open Universe*, p. 44.

counts as a more reasonable hypothesis or general direction of inquiry, changes with, and thus depends upon, the specific details of our context of knowledge.[10]

In a somewhat similar way, recent critics of super-symmetry, string theory and their developments, theories aiming for a unified quantum mechanical approach to the four known forces including gravitation, have become increasingly strident in recent years, pointing to a persistent vagueness which has yielded no significant predictions over a period of some 20 years.[11] The contemporary conflicts and alternatives chiefly turn on varieties of string theory, based in particle physics, and versions of quantum gravity which require no background metric and are more closely related to relativity. Lacking a logic of abduction, we, and the physicists, do best to listen to many voices.

The unification of three of the four known forces as a quantum field theory now promises further developments from that direction. The genius of Einstein remains beyond doubt, of course; and in fact, Einstein foresaw the general conflict. As early as 1916 he wrote that "because of the intra-atomic movement of electrons, the atom must radiate not only electromagnetic but also gravitational energy, if only in minute amounts," thus "…it appears that the quantum theory must modify not only Maxwell's electrodynamics but also the new theory of gravitation."[12] There has been some recognition of the tensions between quantum mechanics and Einstein's physics of space, time and gravitation from the very start.

In the history of theory in quantum gravity, in view of the lack of access to the relevant energies and related empirical evidence, much of the development has been by way of analogies with developments in related fields of study. Quantum mechanical considerations, it was thought, would have to modify Einstein's theory of the gravitational field, and the theory of the gravitational field would have its impact upon quantum mechanics. But how exactly? Summarizing some of the history, Dean Rickles, in a recent textbook on the philosophy of physics, sees the development as a matter of arguments by analogy:

> When quantum gravity was finally studied in a systematic way, it was undertaken to a considerable degree on the basis of analogies with other fields. Inferences were made (not always soundly) on the basis of these analogies to physics at the Planck scale. Later work revealed the inadequacies in this analogical reasoning.[13]

It is sometimes held that arguments from analogy are arguments from the character of one particular to that of another similar particular, and that would seem to

[10] See the opening discussion of the theoretical conflict in the middle of the last century in Stachel, John (1999) [29] "Comments" in Cao, Tian Yu, ed. (1999) [4], pp. 233–234.

[11] The related criticisms are forcefully summarized in Woit, Peter (2007) [31] *Not Even Wrong, The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics*. See also Smolin, Lee (2006) [27] *The Trouble with Physics*.

[12] Einstein, Albert (1916) [5] "Approximate integration of the field equations of gravitation," in Engel, A. ed. (1997) *The Collected Papers of Albert Einstein*, Vol. 6., p. 209. Quoted in Rickles, Dean (2008) [25] *Companion to Contemporary Philosophy of Physics*, p. 285, n. 39.

[13] Rickles, Dean (2008) [25] *Companion to Contemporary Philosophy of Physics,* p. 284.

introduce a significant basis for contrast between arguments from analogy and the presuppositions of abductive inferences.[14] But on the other hand, the analogies involved in quantum gravity appear to be analogies between better established theories and more speculative theories—or conjectural models possibly contributing to such speculative theories. So, the idea might be, for instance, that the gravitational force must be like the electromagnetic force, both involve fields and attractions of one thing to another, and in consequence, since we recognize the photon, with its wave/particle duality, as carrying the electromagnetic force, it seems reasonable to suppose that there must be a similar entity, the graviton, with its own wave/particle duality, to carry the gravitational force.

That this is part of a significant argument from analogy in the history of quantum gravity is evident from that fact that it has been responded to by use of arguments based on important disanalogies, e.g., the argument put forward by the Soviet physicist M. P. Bronstein in the 1930s and later brought to the attention of the wider world.[15] The chief idea arising from related developments is that quantum gravity cannot be formulated by direct analogy with quantum electrodynamics but instead that the unique features of gravitation require special treatment in which some generalization of quantum field theory is needed—one which is applicable in the absence of a fixed background metric. This would be a version of quantum field theory based on a different analogy—an analogy to the dependence of the space-time metric on mass-energy content as in Einstein's physics.

I am sure that many similar illustrations could be provided of projections or extrapolations regarding quantum gravity, and this thought is encouraged by viewing abductive inference as guided by existing theoretical/conceptual regularities. This is to say, in effect, that knowledge of the relevancy, or weight, to be assigned to particular similarities is not known *a priori*, but arises instead from the actual establishment of theory and related habits or patterns of inference—based in the specifics of the theories in question. Still, advances in knowledge can establish new relevancies and disrupt the old, and in consequence, though we are never without some sense for the difference between strong and weak analogies, this is bound to evolve in degree with the growth of knowledge. Such context dependency of the plausibility of analogical and abductive reasoning helps us understand a famous quotation illustrating misplaced early confidence from the likes of Werner Heisenberg and Wolfgang Pauli—in their first 1929 paper on quantum electrodynamics:

[14] Contrast the discussion in Thagard, Paul and C. Shelly (2001) [30] "Emotional Analogies and Analogical Inference," in Gentner, Dedre et al. (eds.) (2001) [8] *The Analogical Mind*, pp. 336–337.

[15] See the discussion of Bronstein's work in Stachel, John (1999) [28] "Introduction: Quantum field theory and space-time," in Cao, Tian Yu, ed. (1999) [4] *Conceptual Foundations of Quantum Field Theory*, pp. 171–173.

> Quantization of the gravitational field, which appears to be necessary for physical reasons, may be carried out without any new difficulties by means of a formalism fully analogous to that applied here.[16]

Obviously, the analogy was not without its appeal, given what they knew or thought most significant at the time; but the present context of physical knowledge renders it much less plausible. Standard quantum mechanics presupposes a fixed background metric—a scheme of locations and times at which particles interact—, while the metric is internal to the gravitational field.

Although it may do some damage to the Peircean scheme of philosophical triads,[17] to see abductive inference as an abstract element of arguments from analogy, this is suggested by emphasis on the role of theoretical context in inferences to explanatory hypotheses. On somewhat similar grounds, comparative evaluation of competing (untested) hypotheses, might be viewed as a matter of stronger and weaker analogies—i.e., comparative evaluation of alternative models which arise from reinterpretation of various living theories. In particular cases, or perspectives of discussion, a moment of abductive inference may stand out, while in other cases some embedding analogy stands in the forefront of attention.

3 Penrose on Quantum Gravity

On of the most fascinating recent proposals regarding quantum gravity is the idea from Roger Penrose that quantum state reduction is a *gravitational* phenomenon. "I belong to the general school of thought that maintains," he writes, "that the phenomenon of quantum state reduction is a *gravitational* phenomenon," and moreover, as he further puts his claims, "essential changes are needed in the framework of quantum mechanics in order that its principles can be adequately married with the principles of Einstein's general relativity."[18]

One way to understand this proposal is to ask why we do not see macroscopic superpositions. If quantum mechanics is a general description of physical reality, which we suppose it is, and subatomic particles and related things (at least up to the size of 60-atom, carbon "Buckyballs") can be placed into quantum mechanical superpositions (the state of being in more than one place at the same time) then this creates the expectation (often dampened in various interpretations of the

[16] Heisenberg, Werner and Wolfgang Pauli (1929) [13] "*Zur Quantenelektrodynamik der Wellenfelder*," *Zeitschrift für Physik*, 56, pp. 1–61. Quoted in Stachel (1999) [28] "Introduction," p. 168.

[17] See my discussion of Peirce's triadic category theory in Callaway, H.G. (2008) [3] "A Role for Peirce's Categories?" in Callaway, H.G. (2008) *Meaning without Analyticity, Essays on Logic, Language and Meaning,* pp. 183–192.

[18] See, e.g., Penrose, Roger and P. Mercer (1998) [20] "Quantum Computation, Entanglement and State Reduction," *Philosophical Transactions, Mathematical, Physical and Engineering Sciences*, Vol. 356, No. 1743, p. 1932.

formalism) that the objects of every-day life should be capable of entering into superpositions—though this is not observed.

One way to envisage the phenomenon of superposition is to consider the classical double-slit experiments. We set up a light source, a laser for example, and shine it in the direction of a screen which will brighten when the light encounters it, and between the screen and the source, an opaque barrier is placed which has two parallel slits through which the light may pass. If either slit is used alone, by covering the other, then the result is a lighted column behind that slit. However, if both slits are opened, then the result is a broad interference pattern of darker and brighter bars across the screen. When only one slit is available, the light seems to travel as discrete particles, coming out more or less directly behind the open slit. However, when both slits are available, then the light behaves in a wave-like fashion. What comes through each slit interferes with what comes through the other—producing a characteristic pattern of dark and light bands. The results sometimes suggest particles, sometimes waves. These results persist, moreover, if electrons are used instead of light, or even if molecules are used; and they persists if the photons or electrons or molecules are sent out one by one. It is as though each particle is in two places at once (in superposition), goes through both slits and then interferes with itself. If the two slits are open, though the particles are sent through one at a time, then the interference pattern slowly accumulates on the screen, while if only one slit is open, then all seems to be a matter of particles going through the one open slit. Even more curious, if a detector is installed to determine which of the two slits a particle traverses, then the interference pattern disappears: collapse of the wave function.[19]

On the assumption that quantum state reduction, the collapse of the wave function, as in measurement, is a *gravitational* phenomenon, the absence of observed macroscopic objects in superposition would seem to make good sense. The greater the mass of an object placed in superposition, the *less likely* it is that such a superposition will be stable for any length of time—superpositions of increasing mass should be increasingly rare. We have an inference to a hypothesis.

Penrose aims to avoid the Copenhagen and "many worlds" interpretations and offer a realist conception. State reduction, or the collapse of the quantum mechanical wavefunction, is a result of gravitational interactions in nature, requiring no observer—it is a result of physical interaction. On his view, gravity pulls objects back into a single location without need of an observer (or multiple worlds), and the more massive an object in superposition, then the shorter the time of any stable superposition.

Penrose is also much concerned to emphasize a methodological conservatism: "the complete agreement that the standard quantum formalism has with all experiments to date;" it is only in newly proposed experimental situations that Penrose's proposal could be verified or experimentally disconfirmed. He also

[19] Compare the discussion in Penrose (2004) [19] *The Road to Reality, A Complete Guide to the Laws of the Universe*, pp. 504–505.

builds on “a certain already existing conflict between the fundamental principles of general relativity and of quantum mechanics.”[20] The conflict is implicit in the “measurement problem” of quantum mechanics “which is to comprehend how, upon measurement of a quantum system, this (seemingly) discontinuous “R-process” [reduction of the state vector or collapse of the wavefunction] can come about”—given the expectations connected with the “U-process,” the uniform linear evolution of an encompassing quantum system “solely according to the Schrödinger equation.”[21]

In order to preserve the expected uniform evolution called for by the Schrödinger equation, some quantum physicists have gone so far as to suppose that multiple, unobservable worlds are required, parallel to ours without subsequent observable interaction—in which, collectively, everything that can happen in accordance with the equation does happen. The Copenhagen interpretation, in contrast, supposes that it is observation which brings about state reduction and a measurable result or outcome.

In contrast to earlier interpretations or approaches to quantum mechanics, Penrose argues for reduction as an objective phenomenon, the wavefunction is a physical wave, and that “present-day quantum mechanics is a limiting case of some more unified scheme, whereby the U and R procedures are *both* to be approximations to some new theory of physical reality.”[22] Though he does not propose anything like a full theory of state reduction, he offers a model which is intended to constrain any broader theory. This builds from a “basic conflict” which Penrose sees between “Einstein’s covariance principle and the basic principles of quantum theory, as they related to stationary states of superposed gravitational fields.”[23]

Einstein’s covariance principle tells us that the forms of physical laws are invariant under arbitrary transformations of coordinate systems. There is no privileged coordinate system. On the other hand, Penrose asks us to imagine a situation in which a very small lump of some rigid material has been placed in a quantum superposition of two positions —this thought experiment is to be regarded as a “an inanimate version of ‘Schrödinger’s cat’.”[24] The argument is that, ignoring gravitation, “the two alternative locations of the lump will each be stationary states,” and in consequence of quantum mechanical considerations arising from the Schrödinger equation, the linear combination of the two is also a stationary state. The overall configuration does not evolve in time.

[20] Penrose, Roger and P. Mercer (1998) [20] “Quantum Computation, Entanglement and State Reduction,” pp. 1932–1933.

[21] Penrose, Roger (1996) [18] “On Gravity’s Role in Quantum State Reduction,” *General Relativity and Gravitation*, Vol. 28, No. 5, pp. 581–582.

[22] *Ibid*., p. 583.

[23] *Ibid*., p. 584.

[24] *Ibid*., pp. 584–585.

Next we are to consider the gravitational fields associated with the two superposed positions of the rigid lump of material. There is also a superposition of the associated gravitational field, and the reader begins to see how Penrose intends to argue for a "gravitational role in state-vector reduction."[25] "But the principle of general covariance denies any significance to particular coordinate systems," he argues, and hence "it asserts that there should be no preferred pointwise identification between two different spacetimes."[26] This is a problem, because quantum field theory, in its usual forms, assumes that a background metric is given, and this assumption is needed to make calculations about possible outcomes. Something else is required.

Given a role for gravity in reduction of the superposition, then this must be in one or the other direction, though neither is privileged. That is the conflict or tension which Penrose envisages. It is partly a conflict between the expectations of a stationary state created by quantum mechanics, in light of the prospect of gravitational interactions, and partly a matter of the fact that there is uncertainty of location related to the non-privileged alternative reductions. It takes energy to sustain the superposed fields (which are slight distortions of space-time generated by the superposition of the associated mass), and the larger the masses involved the more energy it takes. Penrose's proposed solution is to say that the supposedly stationary state of superposition is in reality somewhat like an unstable atom or particle which has a probability of decay to be calculated by consideration of its mass-energy uncertainty. As he elsewhere puts the point, there is a "fundamental energy uncertainty, EG" of the superposition, and,

> The next step is to invoke a form of Heisenberg's uncertainty principle (the time/energy uncertainty relation)… . It is a familiar fact, in the study of unstable particles or unstable nuclei (such as Uranium U_{238}) that the average *lifetime* T, having an inbuilt time uncertainty, is reciprocally related to an energy uncertainty, given by ħ/2T. Now we are going to think of our superposed state $|\Psi> \ = w|\chi> \ + z|\varphi>$ as being analogous to this, itself being unstable, with a lifetime T_G that is related, by Heisenberg's formula, to the fundamental energy uncertainty E_G [of the superposition]. According to this picture, any superposition like $|\Psi>$ would therefore decay into one or the other constituent states,$w|\chi>$ or $z|\varphi>$, in an average time scale of $T_G \approx \hbar/E_G$.[27]

The superposition, $|\Psi>$, though stable, a standing wave, in accordance with the Schrödinger equation, is said to be *analogous* to an unstable nucleus or subatomic particle. The average lifetime, or rate of decay, is then proportional to (a function) ħ, of the Planck constant, divided by the energy uncertainty of the superposition of the

[25] *Ibid.*, p. 583.

[26] Penrose, Roger (2004) [19] *The Road to Reality*, p. 850.

[27] *Ibid.*, p. 853; cf. Rovelli, Carlo (1999) [26] "'Localization' in Quantum Field Theory," in Cao, Tian Yu, ed. (1999) [4], p. 221. Calling for a "theory of quantum geometry," Rovelli argues that given "the need to accommodate the superposition principle, and thus the possibility of quantum superposition of distinct geometries," there can be no "well defined metric structure" for "all quantum states of the theory." This is to reject simple presupposition of a fixed background metric in a fully adequate quantum field theory.

gravitational field—which is a function of the mass of the object placed in superposition.

Since both ħ, a function of the Plank constant, and $E_{G,}$ the energy uncertainty of the superposition, are very small, dividing through these two very small quantities, the average lifetime of the superposition, T, is expected to be measurable under plausible experimental conditions. In this way, Penrose's proposal escapes the constraint of extremely high energies otherwise taken to be required to probe the effects of quantum gravity.

In a somewhat earlier formulation, Penrose writes,

> To compute the decay time, according to this proposed scheme, consider the energy *E* that it would cost to pull away one instance of the mass, moving it out away from coincidence, in the gravitational field of the other, until the two mass locations provide the mass superposition under consideration. I propose that the time scale of the collapse of the state vector of this superposition is of the order of
>
> $$T \sim \hbar/E$$

For a nucleon, this would be nearly 10^8 years, so the instability would not be seen in existing experiments. However, for a speck of water of 10^{-5} cm in size, the collapse would take about 2 hours. If the speck were 10^{-4} cm, the collapse would take about 1/10 sec, whereas for 10^{-3} cm size, the collapse of the state vector would take place in only some 10^{-6} sec.[28]

Again, a superposition may be compared to an unstable nucleus, and the greater the mass of the object in superposition, the greater its instability and the shorter its average lifetime T—as with a transuranic nucleus of great atomic weight. It is clearer in this passage, why the phenomenon in question would not yet have been observed.

Though we have been viewing what is going on in Penrose's proposal as an inference to a hypothesis, in extended stretches of this story, namely the hypothesis that gravitation has a role in state reduction, it appears in the wider context that this hypothesis arises from a certain analogy. This is an analogy between well established quantum mechanical theories of the decay of unstable nuclei and state reduction—functioning as a constraint on the unification of quantum mechanics with general relativity. Part of the appeal of this reasoning arises, because the promise of constraint on possible unification is rare. Though we are dealing with an as yet untested model,[29] intended to constrain some envisaged, but not yet formulated theory, the reasoning clearly enters into a small charmed circle of similar efforts, including, Stephen Hawking on black holes and "Hawking radiation"—significantly bridging the gap between quantum mechanics and general relativity. The supposition of Hawking radiation also arises as a gravitational

[28] See Penrose in Hawking, Stephen and Roger Penrose (1996) [12] *The Nature of Space and Time*, p. 72.

[29] See the plan for an experimental design to test Penrose's hypothesis in Marshall, W.C., Simon, R., Penrose, R., et al. (2002) [15] "Towards quantum superposition of a mirror," *Quantph*/ 0210001v1, pp. 1–5.

effect, since pairs of virtual particles are separated—at the event horizon of black holes. It is in this way we can understand how black holes radiate.[30]

That we find an analogy at the heart of Penrose's proposal suggest a significant dependence of inference to a hypothesis upon analogical reasoning—as an intermediate context supportive of such inference. This creates the expectation that particular abductive inferences should share the strength or weakness of corresponding analogies.

4 The Virtues of Hypotheses

One way to get an overview of the virtues of hypotheses is to see them as spanning the ever-present gap between the universal aspiration and particular established theory and facts. We want a new hypothesis and a theory with great comprehension so that, if it is correct, it will have a maximum tendency to avoid future disappointments or disconfirmation. That is one of our methodological ideals. Still, we have contrasting ideals. Though we want generality, we also want testability, and connected with this is the ideal of preserving as much as possible of accepted theory, even as we go about changing it in light of contrary evidence.

No one would have even considered Einstein's physics, without the assurance that it came up with the same predictions as Newtonian physics over the very wide range of circumstances in which Newtonian physics had succeeded in its predictions—so that even the boldest of hypotheses must have its conservative side. But, on the other hand, holding that every bold hypothesis must have its conservative side, does not plausibly amount to saying that new theories cannot be "revolutionary."[31] A new theory might plausibly be regarded as "revolutionary," if it takes in new predictions, and preserves the evidence supporting its older competitor, while significantly modifying the principles or laws which allowed the comprehension of supporting evidence by the older competitor.

Seeing the virtues of hypotheses as spanning the tensions between the particularity of established fact and theory and ideal universality, as aiming us toward both predictive tests and general explanatory intelligibility, the other virtues fall somewhere between, in somewhat the following order: Refutability, conservatism, modesty, precision, elegance, generality. I want to suggest a continuum of the virtues with contrasting extreme points, approximately, from the virtues of the experimentalist to the virtues of the theoretician.

In Table 1, each has been outfitted with familiar, named excesses and defects, to help with their recognition in Aristotelian style, and the virtue of "*simplicity*" is understood as a component of both "*modesty*"—when accepted theory is chiefly retained—and of "*elegance*"—as involved in relating evidence and prediction to

[30] See, e.g., Hawking, Stephen (1988) [10] *A Brief History of Time*, pp. 105ff.

[31] Contrast Barrow, John D. (2003) [2] *The Constants of Nature*, p. 60.

Table 1 Virtues of hypotheses

(Particular) ←----------------------------/ / /--------------------------------→ (Universal)

Excess:	implausibility	dogmatism	meekness	*Übergenauigkeit*[a]	naiveté	rigidity
Virtue:	**Refutability**	**Conservatism**	**Modesty**	**Precision**	**Elegance**	**Generality**
Defect:	self-insulation	extravagance	vanity	vagueness	complexity	bias

(Experimentalist) ←--------------------------/ / /-------------------------→ (Theoretician)

[a] "*Übergenauigkeit*," German: over-exactness; compare Latin, *meticulosus*: fearful

broader innovations of theory. (We tend to speak of "*elegance*" when a broader range of poignant evidence is comprehended by simplicity of law or principle.) The virtues of *refutability*, *conservatism* and *modesty* appeal to the experimentalist, because they involve only limited modifications of accepted theory, and because accepted theory tends to be built into the instruments and methods which have been used in testing and developing accepted theory. The further the theoretician departs from accepted theory, the more likely it is, in general terms, that innovative hypothesis and theory will make no clear predictions, even though they agree with all evidence so far established. In this direction science tends to the speculative. Still, it is not impossible for a very innovative theory to come up with plausible predictions of otherwise unforeseen phenomena, and a growing number of problems in accepted theory may bring the theoretician to the conviction that modest tinkering has degenerated into meekness, so that an entirely new approach to outstanding problems is required. *Precision* and *elegance* are particularly important in any broadly innovative approach, because *precision* (mathematical precision and quantification in particular) makes it more reasonable to expect a range of measurable results; and *elegance* holds the promise of a wealth of evidence comprehended on the basis of (relatively) simple principles—all of which would take us in the direction of greater *generality* than what was heretofore established.

If we follow Peirce in holding that "Logic may be defined as the science of the *laws* of the stable establishment of beliefs,"[32] then we may doubt that there is, or could be, a *logic* of abduction which could establish stable beliefs regarding untested explanatory proposals on the basis of *laws*. Our list of the virtues of explanatory hypotheses is not a matter of general laws, since we have no general means of ranking the comparative importance of, say, conservatism and modesty or simplicity, precision or generality for an arbitrarily selected context of inquiry. In consequence, it seems we have no formal or law-like way to arrive at a generalized ranking of competing hypotheses which exhibit these virtues. The named virtues are

[32] See Peirce, 1896 [16] [*Collected Papers,* Vol. 3, paragraph 429], where he continues by defining "exact logic": "Then, exact logic will be that doctrine of the conditions of establishment of stable belief which rests upon perfectly undoubted observations and upon mathematical, that is, upon diagrammatical, or, iconic, thought."

comparative terms. One hypothesis is judged to be more easily refutable or more conservative, more modest, simpler, more precise or more general than another hypothesis in relation to the same domain and context of inquiry, and the non-comparative, presumably monadic predicates, “refutable,” “conservative,” modest” “simple,” etc. borrow what sense they have from the specific comparisons made in specific contexts. But even in a particular context of inquiry, when we are dealing with a specific domain and accepted theory and its problems, if we know exactly which of the proposed new explanatory hypotheses can be justly called more easily refutable, more conservative, simpler, etc., this alone does not tell us which hypothesis might best be accepted for preferential empirical examination in that context of inquiry. It is not that such judgments are not in fact made, and it is not that we cannot see the wisdom of examples. But our assembly of convincing examples of preferences among the virtues, from the contexts of particular inquiries, do not add up to a *logic*, as contrasted with an art, of abduction. In the context of possible theory change, we no longer know what to count as purely formal elements, as is perhaps evident, say, from Einstein’s revision of Newton’s definition of force, or from his revision of the concept of simultaneity.

On some occasions, generality overrules simplicity or conservatism, on other occasions modesty or conservatism rightly trumps generality or elegance. But if there were truly a *logic* of abduction, then we would expect general rules or some stable ordering of the virtues across distinctive domains, contexts and occasions of inquiry. Our esteem for any particular virtue in the order it provides to a range of hypotheses on a particular occasion appears to be bound to the particular context, and dependent on the specifics of content, and it is otherwise chiefly retrospective and *ex post facto*. No generalized ordering of explanatory hypotheses in terms of the virtues seems to be projectable across all domains and occasions of inquiry: not conservatism, not boldness.

This is not to say, however, that our intuitive sense for the value of particular analogies, and related hypotheses, cannot be improved by means of the study of the mathematics of model theory—which represents the abstract possibilities of projective or analogical mappings. A chief point, however, is that knowledge of the mathematical possibilities of mappings is empty without the knowledge of particularities of the various domains—which helps establish a required sense of relevancy and salience. Corresponding model-theoretic analogies are of general interest in supporting abductive inference, because model-theoretic analogies are intuitive or natural—at least for those who have some understanding of model theory—but also, in part, on the assumption that mathematics generally can be paraphrased into set theory. We might imagine, of course, that physicists, and other scientists, sometimes craft analogies supportive of particular hypotheses formulated in highly sophisticated mathematical terms, making no direct suggestion of model theory, strictly considered. But even in such cases, the proposals ought to be open to paraphrase into model-theoretic analogies.

Notice that physics seems lately to have discounted modesty and even refutability as it has explored the luxurious mathematical possibilities of string theory, drawn in this direction by the prospect of a unified quantum-mechanical theory of the four known fundamental forces—a prospective theory which would bridge and reconcile quantum mechanics and general relativity. General relativity must break down, Stephen Hawking has argued, where the erstwhile continuities of curved space-time reach quantum-mechanical levels forbidding exact continuity of competing measurements.[33] Even at the time of Einstein's early work, though it was known that Newtonian physics didn't correctly predict the orbits of electrons around the atomic nucleus, the theory of relativity made only small and inadequate correction to these faulty predictions. In this context we understand the significance of Einstein's role as the creator of relativity theory *and* as one of the chief thinkers responsible for the birth of quantum mechanics.[34] We see more clearly now, or believe more firmly, that continuity of space, time and motion must, in some fashion, give way to quantum-mechanical indeterminacies.

5 The Scientific Imagination

My conclusion is that in selecting among untested hypotheses, we have to do with a highly contextual type of judgment, a kind of art or wisdom arising from the expert's extensive familiarity with the subject-matter. Selecting plausible hypotheses and weighing of analogies is not a formalizable skill so much as it is a matter arising from flexibility of mind in encompassing the details of a subject-matter. We have also found some reason to contrast the familiarity of the theoretician with that of the experimentalist. Deep familiarity with the details and problems of the domain of inquiry, on the part of a master of the discipline, is the one commonality which bridges those cases where we are inclined to favor bold generality and those where we may be inclined to favor more conservative, modest or easily testable alternatives. It is from this kind of perspective that we may judge of the lack of relevant detail, or the amateur status of wild guessing—as contrasted with educated guesswork.

The scientific imagination uses an organic classification, we might say, understood as a matter of strong, detailed analogies, while amateur fancy joins by accidental resemblances. The distinction parallels that between a more compelling argument from analogy and the kind of weak or false analogy which ignores detailed differences to focus on superficial similarities. In knowing what to count as potentially useful or more useful theoretical innovation and what might count as vain fancy, we depend on detailed reference to the particular subject-matter in the

[33] Cf. Hawking, Stephen (2006) [11] *A Briefer History of Time*: "…we know that the theory of general relativity must be modified. Because the classical (i.e. non quantum-mechanical) version predicts points of infinite density—singularities—it prognosticates its own failure…" See, p. 119 in the German edition.

[34] Einstein received the 1921 Nobel Prize in Physics, for his photoelectric law and work in the field of theoretical physics, thus chiefly for his work in the origin of quantum theory. Relativity, still under debate, was not mentioned.

continuity of inquiry. In initial evaluation of hypotheses proposed, we must start from a detailed and systematic account of past accomplishments in a field, together with the outstanding anomalies and problems. It is out of this tension that the properly disciplined and genuinely creative scientific imagination arises.

References

1. Aliseda, A.: Abductive Reasoning, Logical Investigations into Discovery and Explanation. Springer, Berlin (2006)
2. Barrow, J.D.: The Constants of Nature. Vintage Books, London (2003)
3. Callaway, H.G.: A role for Peirce's categories? In: Callaway, H.G. (ed.) Meaning without Analyticity, Essays on Logic, Language and Meaning, pp. 183–192. Cambridge Scholars Publishing, New Castle (2008)
4. Cao, T.Y. (ed.): Conceptual Foundations of Quantum Filed Theory. Cambridge University Press, Cambridge (1999)
5. Einstein, A.: Approximate integration of the field equations of gravitation (1916). Translated in Engel, A. (ed.) The Collected Papers of Albert Einstein, vol. 6. Princeton University Press, Princeton (1997)
6. Feynman, R.: QED: The Strange Theory of Light and Matter. Princeton University Press, Princeton (1985)
7. Gabbay, D.M., Woods, J.: The Reach of Abduction, A Practical Logic of Cognitive Systems, 2 vols. North-Holland, Amsterdam (2005)
8. Gentner, D., Holyoak, K., et al. (eds.): The Analogical Mind, Perspectives from Cognitive Science. MIT Press, Cambridge (2001)
9. Greene, B.: The Elegant Universe. Vintage Books, London (1999)
10. Hawking, S.: A Brief History of Time, from the Big Bang to Black Holes. Bantam Books, New York (1988)
11. Hawking, S.: A Briefer History of Time, the German edition. Rowohlt, Hamburg (2006)
12. Hawking, S., Penrose, R.: The Nature of Space and Time. Princeton University Press, Princeton (1996)
13. Heisenberg, W., Pauli, W.: Zur Quantenelektrodynamik der Wellenfelder. Zeitschrift für Physik **56**, 1–61 (1929)
14. Magnani, L.: Abductive Cognition, The Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Heidelberg (2009)
15. Marshall, W.C., Simon, R., Penrose, R., et al.: Towards quantum superposition of a mirror. *Quant-ph*/0210001v1, Sept. 2002, pp. 1–5
16. Peirce, C.S.: The Collected Papers of Charles Sanders Peirce, vol. 3. Harvard University Press, Cambridge (1931–1935)
17. Peirce, C.S.: Pragmatism as the logic of abduction (1903). In: Houser, N., Kloesel, C. (eds.) The Essential Peirce, Selected Philosophical Writings, vol. 2 (1893–1913). Indiana University Press, Indianapolis (1998), pp. 226–241
18. Penrose, R.: On Gravity's Role in Quantum State Reduction. Gen. Relat. Gravit. **28**(5), 581–582 (1996)
19. Penrose, R.: The Road to Reality, A Complete Guide to the Laws of the Universe. Vintage Books, New York (2004)
20. Penrose, R., Mercer, P.: Quantum Computation, Entanglement and State Reduction. Philos. Trans. Math. Phys. Eng. Sci. **356**(1743), 1927–1939 (1998)
21. Popper, K.: The Poverty of Historicism. Ark, London (1957)
22. Popper, K.: The Open Universe. Routledge, London (1982)
23. Popper, K.: Objective Knowledge, revised edition. Oxford University Press, Oxford (1979)

24. Quine, W.V., Ullian, J.S.: The Web of Belief, 2nd ed. Random House, New York (1978)
25. Rickles, D.: Companion to Contemporary Philosophy of Physics. Ashgate, Burlington (2008)
26. Rovelli, C.: 'Localization' in quantum field theory. In: Cao, T.Y. (ed.) Conceptual Foundations of Quantum Field Theory, pp. 207–232. Cambridge University Press, Cambridge (1999)
27. Smolin, L.: The Trouble with Physics. Houghton Mifflin, Boston (2006)
28. Stachel, J.: Introduction: quantum field theory and space-time (1999). In Cao, T.Y. (ed.) Conceptual Foundations of Quantum Field Theory, pp. 171–173. Cambridge University Press, Cambridge (1999)
29. Stachel, J.: Comments. In: Cao, T.Y. (ed.) Conceptual Foundations of Quantum Field Theory. Cambridge University Press, Cambridge (1999)
30. Thagard, P., Shelley, C.: Emotional analogies and analogical inference (2001). In: Gentner, D., et al. (eds.) The Analogical Mind. MIT Press, Cambridge (2001)
31. Woit, P.: Not Even Wrong, The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics. Random House, London (2007)

Galileo's Discovery of the Moons Orbiting Jupiter Based on Abductive Inference Strategies

Jun-Young Oh and Yoo Shin Kim

Abstract The objective of this study is to understand the scientific inferential processes of Galileo's discovery of Jupiter's moons. Abductive reasoning has played very important roles in producing creative leaps and breakthrough for scientific discovery in history of science. This article presents a scientific procedure that involves abductive inference in general. And we propose a noble and refined model for abduction inference and show its validity by applying to the inferential process of "Galileo's discovery of the moons of Jupiter", with historically considered evidence. It makes three broad macro perspectives; rather than only hypothetico-deductive method, (1) "fixed stars hypothesis suspected", (2) Moon hypothesis can be suggested and selected by abductive strategies, (3) Moon hypothesis expansion.

1 Introduction

Galileo's work has been commonly cited in scientific textbooks as a prime example of an empiricist methodology, and these views of the empiricist tradition have become widely accepted. According to [35, p. 4], Whewell and Mach viewed Galileo as a collector of facts who discounted Aristotle's search for causes and instead sought to employ inductive methods. As Francis Bacon noted, induction by enumeration is an overly simple method, and the hypothetico-deductive method is regarded as a great improvement. With the rise of logical positivism, the

J.-Y. Oh (✉)
Hanyang University, Seoul 133-791, Republic of Korea
e-mail: jyoh3324@hanyang.ac.kr

Y. S. Kim
Science Studies, Pusan National University, Pusan 609-735, Republic of Korea
e-mail: kimys@pusan.ac.kr

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_16, © Springer-Verlag Berlin Heidelberg 2014

methodology in Galileo's work has gradually assumed canonical form as the hypothetico-deductive (HD) method. But hypothetico-deductive (HD) method as scientific reasoning does not always account for breakthroughs for scientific discoveries in history of science.

Hanson [4] argued that:

> Kepler did not begin with the hypothesis that the orbit of Mars around the sun was elliptical and he argued that Kepler deduced his conclusions confirmed by Tycho Brahe's observational data through hypothetico-deductive reasoning. ······ Instead, Hanson suggested that Brahe's unexplained observational results revealed the problem and became Kepler's starting point. He struggled to overcome, to one hypothesis then to another before postulating his hypothesis of the elliptical orbit (Hanson, pp. 72–73).

> Kepler did not begin with the hypothesis that the orbit of Mars around the sun was elliptical and he argued that Kepler deduced his conclusions confirmed by Tycho Brahe's observations. These latter observations were given, and they set the problem—they were Johannes Kepler's starting point. He struggled back from these, first to one hypothesis, then to another, then to another, and ultimately to the hypothesis of elliptical orbit (Hanson, p. 72).

Hanson's argument leads us to consider what Peirce calls "abduction or abductive inference". "According to Peirce, abduction is the process of forming an explanatory hypothesis" (5.172), which "must cover all the operations by which theories and conceptions are engendered (5.590), including not only the invention of hypotheses but also selection of them for further consideration" [9, p. 477]. As we know, Galileo has an important position in the history of science. Among his various achievements, astronomical observations made great contribution to history of science. Especially, sunspots and the phase variations of Venus were greatly important discoveries in the history of science. And Galileo studied the moon's surface and Jupiter's moons very carefully, which helped Copernican system be in more stable position over Ptolemy's system.

The strategy of this article is composed of three steps. The first step is to review and discuss Peirce's abduction and Magnani's refined version of abduction.

Second step is to describe key elements of Galileo's observation procedures and find those of the thought pattern that guided Galileo's discovery. Finally, we propose an epistemological model of the scientific process, sophisticated model of abduction. And with this model we interpret Galileo's discovery of the moons orbiting Jupiter. Next section, we discuss the abductive inference as background.

2 Background

Abductive inference

Josephson and Josephson [7] state:

> Abduction, or abductive inference, is a form of inference from which we employs data to develop a theory that best explains those data and results (unlike simple induction by

> enumeration). So abduction is a kind of explanatory or interpretive inference. In view of theory construction, abduction is a type of theory-forming. [7, p. 5].

A pattern was suggested by Peirce in his 1903 lectures on pragmatism. Abductions ought to conform to the following pattern which was suggested by Peirce in his 1903 lectures on pragmatism [4, p. 86]:

[PA] 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. (CP 5.189)

These premises of schema accurately portray scientific procedure because research typically begins with a problem (a puzzling or surprising phenomenon) and seeks to derive a solution for an explanatory hypothesis [9, p. 480]. In this scheme, abduction can be understood as a mode of inference that seeks explanations for anomalous or surprising phenomena. Here, A might be of various forms, an hypothesis, or a general theory (or, using the earlier terminology, it might include both a "rule" and a "case"). In this abduction scheme, what Peirce claim is that A is not the conclusion, in and of itself but is rather the assertion that we have reason to suspect that A is true. This means that A is a conjecture for starting point to search for true correct theory or hypothesis. Therefore abduction involves other inference processes.

There are often several possible abductive explanations to explain a fact, but only one (or a few) that provides as the best explanation. There is only one seen as best explaining the event, namely, the one that really happened. However, other preference criteria may be appropriated too, especially when we have no direct test available. Thus, abduction is connected to both hypothesis construction and hypothesis selection [1, p. 33].

Peirce stated the following:

1. Every inquiry whatsoever takes its rise in the observation of some surprising phenomenon (CP, 6.469).
2. The inquiry begins by pondering these phenomena in all their aspects, in the search for some point of view whence the wonder may be resolved. At length, a conjecture arises that furnishes a possible Explanation, by which I mean a syllogism exhibiting the surprising fact as necessarily consequent upon the circumstance of its occurrence together with the truth of the credible conjecture, as Premises (CP, 6.469). According to [24], abductive reasoning strategies can guide the process and the way premises are searched for. According to this approach, an abductive inference starts from the small details and characteristics, and the goal is to find a hypothesis that would explain these details "as a matter of course ([PA], Second Premises)" (CP 5.189). Clues and minute details give hints and suggestions for hypotheses (CP 2.755). The initial conception of the conditional premise is not the result of an inference but is rather an acceptance of the inference. The guess is a reasoned adoption on probation (MS 692: 36). "The very phrase 'as a matter of course' indicates a degree of intuitiveness, a point underscored by the fact that an explanatory conditional conveys a connection of necessity or high probability" (8.231, 7.36) [8].

3. Based on this explanation, the inquirer is led to regard his conjecture, or hypothesis, with favor. As I phrase it, he provisionally holds it to be "plausible" (CP, 6.469).

Thus, we propose that premise one (1) in [PA] corresponds to the "*surprising observations*". Premise two (2) in [PA] corresponds to the "*conjecture and invention of hypotheses*" and the conclusion (3) in [PA] corresponds to the "*selection of hypotheses*".

Retroduction should be separated from Abduction

How can abduction be a form of inference distinct from deduction and induction (as the unfettered play of amusement or as a response to a surprising fact) and a form of recursive analysis that includes deduction and induction? To adapt to new phenomenon or surprising fact like C, we may conjecture many A's. Which is the most appropriate? We need to retroduct to background fact or theories as recursive analysis. This retroduction can eliminate much of the confusions surrounding abduction.

"The distinction between before-trial and after-trial in the evaluation of hypotheses is made in the classical formulation of the HD method. For example, William Whewell required that a hypothetical theory ought to "explain phenomena which we have observed" and "foretell phenomena which have not yet been observed", and they are "of a kind different from those which were contemplated in the formulation of our hypothesis" [37, p. 62–65].

Imre Lakatos [11] states:

> The philosopher of science must survey the scene to see whether there exist alternative hypothesis that do not imply the evidence. Old evidence provides support only within the context of competition between hypotheses [15, p. 227].

According to [29, p. 8], Peirce saw qualitative induction as an evolutionary process of variation and selection. As we have seen, two component processes are involved. (1) Hypotheses projection or abduction: the purely conjectural proliferation of the entire gamut of alternative explanatory hypotheses those are relatively plausible. (2) Hypothesis testing or *retroduction*: the elimination of inappropriate hypotheses based on observational data.

The process of science proceeds via the repeated elimination of rival hypotheses in favor of one preferred candidate. Each stage in the *abduction-retroduction* cycle reduces a cluster of conjectural hypotheses to an accepted theory.

Paavola [25] states:

> Abduction is the first phase of inquiry that generates ideas. As distinct from the evidential viewpoint, the methodological viewpoint emphasizes abduction as one phase in the process of inquiry. Hypotheses and ideas generated with abduction should be tested with deduction and induction (Originally in [25, p. 132]).

Above all, Wuisman [39] state, "the gap between the appearance of a statue observable through the senses and the underlying mechanism from which it emerges seems unbridgeable. What mode of inference or logic can link the sensory perceptions of the statue to this underlying mechanism? The mental acrobatics requires a creative leap, rather than induction or deduction. In the literature on the various modes of inference, it is found that the only reasoning involved in a creative leap is abduction".

Analogical abduction, or rule- forming abduction, is necessary for the creative leap [22]. In the scientific process, it is necessary to update or recreate existing hypotheses for more plausible hypotheses based on newly available information [26, p. 6]. These changes may be conducted after retroduction as preliminary tests, after induction to corroborate a hypothesis, or as rejected tests.

Magnani's epistemological models of hypothetical reasoning (2001)

It is agreed that abduction is the first phase of inquiry used to generate ideas, and hypotheses and ideas generated via this abduction should then be tested by using deduction and induction.

Magnani [16] has developed:

> With others [18, 28] an epistemological model of medical reasoning, called the Select and Test Model (ST-Model), which can be described in terms of the classical notion of abduction, deduction and induction. It describes the different roles played by these basic inference types in developing various kinds of medical reasoning (diagnosis, therapy planning, and monitoring), but it can be extended and regarded as an illustration of scientific theory change. A hundred years ago, Peirce interpreted abduction as an inferential creative process that is used to generate new hypotheses. There are two main epistemological meanings of the word abduction [17]: abduction may simply generate plausible hypotheses (selective or creative) or make it possible to infer the best explanation and thus to evaluate hypotheses. "Selective" abduction will only produce hypotheses for further examination that have some chance of being the best explanation. Selective abduction will always produce hypotheses that at least provide a partial explanation and therefore have "a small amount of initial plausibility". From a syllogistic viewpoint, if abduction is a means of inferring the best explanation as advocated by Peirce, the final explanation might need to be "the most plausible". This classical understanding of selective abduction as inference of the best explanation is described in Peirce's epistemological model as part of the complete abduction-deduction-induction cycle [16, pp. 220–222].

Thus, selective abduction involves making a preliminary guess that introduces a set of plausible diagnostic hypotheses; deduction is then used to explore their consequences, and induction is used to test them with available data. These tests are employed either (1) to increase the likelihood of a hypothesis's being plausible by noting evidence explained by that hypothesis rather than by competing ones or (2) to refute all but one hypothesis.

If new information emerges during this first cycle, hypotheses that have not been previously considered can be suggested, and a new cycle will begin. In this case, the non-monotonic character of the abductive reasoning process is clear.

Once produced, abductive explanations are public objects of 'justification,' which can be checked and tested by independent logical criteria. What are these criteria? Explanatory criteria are necessary because the rejection of a hypothesis requires one to demonstrate whether a competing hypothesis provides a better explanation. For instance, when we choose scientific hypotheses or theories in which the role of "explanation" is dominant, we can reach conclusions according to rational criteria such as *consilience* or *simplicity* [19, p. 27]. To achieve the best explanation, one must have or establish a set of criteria for evaluating the competing explanatory hypotheses reached by creative or selective abduction [19, p. 26].

According to [27, p. 157] and [20], the Bayesian methodology that advanced the mode of justification developed by [30, p. 118] was considered the prior probability used in discussing the plausibility of hypotheses. It is necessary to consider plausibility, analogy, simplicity, and coherence, not only in justifying hypotheses but also in developing them.

The modified theory must be conservatively and dynamically consilient. The hypothetico-deductive method neglects this dynamic feature of theory evaluation. Until now, we have treated consilience as a property of theories, but generalizations can also be understood as best explanations. In addition, it would appear that the maximally consilient hypothesis or theory is one that explains any fact whatsoever. This will occur if there is sufficient flexibility in the set of auxiliary hypotheses to ensure that any phenomenon can be covered by the theory. The level of simplicity is the level of conceptual complexity of hypotheses when their consilience is equal. Their evaluation in this capacity is strongly influenced by Occam’s razor. Simplicity can be highly relevant to the analysis of competing explanatory hypotheses [19, p. 26].

A new cycle begins if new information suggests the hypotheses that have not been previously considered. The epistemological model stresses is cyclical and non-monotonic in character. For example, new information can decrease the likelihood of an earlier hypothesis to zero [26, p. 125]. Thus, non-monotonic inference is time-dependent logic [34, p. 5]. Magnani’s S-T epistemological model is consisted of abduction, deduction, induction, and recycles, and in individual stage we enhance it in follow chapter.

One of Magnani’s central distinction is the distinction between abduction (1) aiming only at generating (plausible) hypotheses, and (2) abduction considered as the inference to the best explanation (IBE), which also evaluation of hypotheses [19, p. 19]. The IBE formulated by Harman have concentrated more on evaluation and justification of suggested hypotheses rather than emphasizing on generative aspects of hypotheses. But Lipton’s interpretation on IBE has brought it closer to the generative interpretation. Therefore we claim that these IBE are considered as retroduction, because it is weak evaluation, and then practical evaluations of suggested and selected hypotheses are Hypothetico-Deduction (deduction-Induction).

Galileo’s research strategies for abstraction

Galileo insisted on the importance of abstraction and idealization in physics, extending the reach of inductive methodology [15, p.49]. In Galileo’s own work, he used idealizations such as the concepts of a ‘free fall in a vacuum’ and the ‘inertial law’ in no friction. These phrases did not occur in directly visible phenomena. Similarly, Galileo discovered the moons of Jupiter during a time when the role of the planets and moons was under debate. His discovery of Jupiter’s moons is an example of the theory-dependence of observation. Galileo interpreted the fuzzy smudges of light as moons' revolving around other planets according to Copernicus's heliocentric principle, while other observers believed that the blurs had no significance. Because they could not interpret the patterns within the framework of Aristotle's search for causes, they assumed that the telescope was unable to display clear images of cosmic bodies.

Archer [2] interpret abstraction according to its traditional meaning of focusing upon certain aspects of something to the (momentary) neglect of others. It is a process of focusing on some feature(s) of something(s) while others remain in the background (p. 170).

Aspects of Bhaskar's interpretation will be discussed using the concept of the theory-dependence of observation. This theory-dependence of observation put Galileo in conflict with the Aristotelian geocentric system because Galileo did not want to accept the simple hypothesis that the earth was positioned at rest at the center of the model. According to Bhaskar, this way of thinking was hypothetico-deductive because it was based on a causal connection. However, McMullin preferred to use Peirce's concept of abduction and considered it retroductive [35, pp. 11–12]. We also argue that the scientific inferences made during Galileo's discovery of the moons orbiting Jupiter required the use of an abductive inference methodology involving *abstraction strategies* associated with the heliocentric hypothesis.

3 Outline of the Scientific Inference Method Based on Abductive Inference Strategies Involving the Deduction-Induction Recursive Cycle [23]

The generation of hypotheses

Data *Abstraction*: A process of *abstraction* or *reconstruction* can be used to organize available data into a small set of necessary and important entities to explain puzzling, surprising, and unfamiliar evidence via theory-dependence of observation. This data abstraction strategies can be covered through all scientific inference process.

Analogical Abduction: Abduction is then used based on an analogical strategy that involves conjecturing and inventing a set of hypotheses based on puzzling and unfamiliar evidence using reconstruction strategies. The hypotheses generated based on analogical abduction are more plausible than competing hypotheses based exclusively on simple abduction. Based on our prior store of declared knowledge in other domains, we use Analogical abduction to invent a hypothesis (a tentative explanation) for a puzzling or surprising phenomenon, based on existing knowledge in other domains [23]. For example,

., an atomic nucleus is observed by Rutherford.

If . Planet (p2) orbit the Sun (p1) in Heliocentric theory (T1), *and atomic nucleus (p4) in atomic theory (T2) is like Sun (p1) in the Heliocentric theory (T1),*

Then, perhaps electrons (p3) will orbit an atomic nucleus (p4).

Therefore (Hence), there is reason to suspect that Rutherford's atomic theory is true.

Retroduction: Once the hypotheses have been developed, they must be ranked and selected by consilience [33] to measure how much each hypothesis explains so that the evaluation phase for the preferred hypothesis can be planned. Thagard's discussion of dynamic consilience [32, p. 84] suggests that he valued data prediction more than data accommodation. According to [29], these hypotheses were tested by exploiting them as a basis for predictions that were then verified. Peirce called this process of eliminating hypotheses via experiential testing (p. 3) and identifying more plausible hypotheses the process of *retroduction*.

When the consilience levels of the hypotheses are equal, evaluating *simplicity* [19, p. 26] makes it possible to address differing levels of conceptual complexity by identifying the more plausible (the simple) of the competing hypotheses.

Updating: Updating and selecting: Hypotheses involve updating or recreating existing hypotheses based on newly available information [26, p. 6] to develop more plausible hypotheses. This stage may be independent of newly available information (data, theories).

The testing of hypotheses

Deduction and Induction Stage: Hypothetico-deduction cycle. Route A or B (Recycle stage)

Deduction: *Deduction* is used to generate additional predictions. This process requires declarative knowledge [14]. The final explanation is "*the most plausible*". The resulting inferred best explanation is described in Magnani's epistemological model as part of the complete abduction-deduction-induction cycle [16, pp. 220–222].

Induction: We then make the necessary observations to match our predictions [14]. The process of *induction* reduces the uncertainty of established hypotheses by comparing the results with observed facts [16, pp. 221], making it possible to identify more plausible hypotheses (dynamic consilience).

Recycling for expansion or revision: We can require that the final explanation should be the most plausible explanation based on the complete abduction-deduction-induction cycle. Induction can be used to identify hypotheses whose expected results are consistent with the observed data and to *updating -deduction-induction (recycles)* to develop more plausible hypotheses (*Route A: Expansion*: the observation outcome corresponds to the expected result).

If new information has emerged (*Route B: monotonic inference*) during the previous cycle, then a new Abduction-Deduction-Induction cycle must be undertaken to achieve more plausible hypotheses (in terms of radical, dynamic and static consilience).

New information can significantly decrease the likelihood of a previous hypothesis, even to zero [26, p. 125], non-monotonic inference is time-dependent logic [34, p. 5].

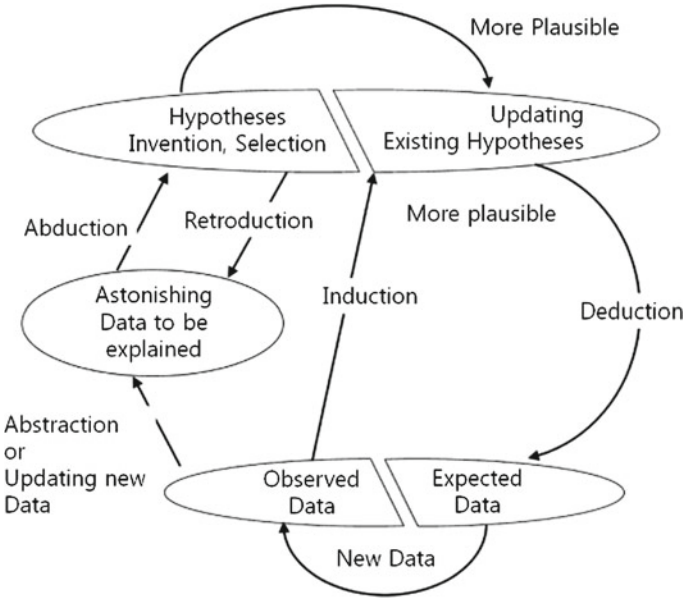

More Plausible
Hypotheses Invention, Selection
Updating Existing Hypotheses
Abduction
Retroduction
More plausible
Astonishing Data to be explained
Induction
Deduction
Abstraction or Updating new Data
Observed Data
Expected Data
New Data

Fig. 1 Scientific inference method based on abductive inferences for the generation of hypotheses involving hypothetico-deduction (Deduction-Induction) method for the evaluation of hypotheses

Our discussions mentioned above about the abductive inferences are portrayed clearly in Fig. 1. We see various cycles in which there are various small cycles Two major cycles and some minor cycles which play roles of connecting the inferences. When you scrutinize these cycles, you can see what the whole process of abductive inferences look like. Through passing the cycles various times, we can reach more plausible hypotheses or theories.

4 Application of our abduction model to Galileo's Discovery of the Moons Orbiting Jupiter

By the end of 1609, Galileo had improved his magnification by 20 times. When he used the instrument to look at the stars, he saw many that were invisible to the unaided human eye—stars that had remained hidden since creation, awaiting discovery [6, p. 125]. Galileo's knowledge of heavenly objects included the three following assumptions [13]:

1. Fixed stars are immovable because they are embedded in an external celestial sphere.

2. Some objects within the celestial sphere—the planets—orbit the sun (e.g., the Earth, Venus, and Jupiter).
3. Other objects—moons—orbit the planets that orbit the sun (e.g., our moon).

Charles S. Peirce had already distinguished among three modes of scientific reasoning in his Harvard lectures during the spring of 1865. Starting from Aristotle's doctrine of induction as a way to infer a syllogism's major premise, Peirce observed that there is "a large class of reasoning" that is neither deductive nor inductive: reasoning a posteriori to a physical hypothesis, or inference of a cause from its effect [21]. Peirce argued that, besides deduction and induction, there is a third mode of inference, which he called "hypothesis" or "abduction". He characterized abduction as reasoning "from effect to cause", and as "the operation of adopting an explanatory hypothesis". Peirce's ideas about abduction, which are also related to earlier historical accounts of heuristic reasoning (the method of analysis), have been seen as providing a logic of scientific discovery. Alternatively, abduction is interpreted as giving reasons for pursuing a hypothesis [21]. The argument's overall logical form is implicit in Aristotle's statement in Posterior Analytics 13, which Zabarrella recognized as an instance of the demonstrative regressus. Following Galileo's characterization of this in D3.3, it involves two progressions, one from effect to cause and the other from cause to effect, separated by an intermediate stage in which one sees the causal connection between the two as necessary and adequate to explain the phenomena [35 , p. 194].

We propose that both the progression "from effect to cause" and the "intermediate stage [35, p. 202]" involve the "*generation of hypotheses*" and that the progression "from cause to effect [35, p. 203]" involves "*the testing of hypotheses*". We also propose that the starting point of the analysis occurs when "*the cause for the effect is materially suspected* [35, p. 202]".

It makes *three broad macro perspectives* about the history of science based on Galileo discovery, first shared ideas about "fixed stars hypothesis suspected", Second, these alternative ideas about "Moon hypothesis" can be suggested and selected, Finally, these idea, Moon hypothesis, are expanded more plausible.

4.1 Fixed Stars Hypothesis Suspected

The cause for the effect is materially suspected as puzzling observations

Effect: noticed three small bright objects close to the giant planet Jupiter.

Cause: The conventional cause for the observation of the effect is due to shining immovable stars. This cause is materially suspected, but not recognized formally as the cause.

The generation of hypotheses

Data Abstraction or Reconstruction: Galileo first observed the moons of Jupiter on January 7, 1610. He had no idea they were moons at the time and simply noticed three small bright objects close to the giant planet Jupiter, which he described in his notes as "stars" [38, p. 108]. it turned out to be a momentous observation. At the time, Jupiter was close to opposition, at its closest approach to earth, and was the brightest object in the evening sky [35, p. 201].

Analogical abduction: (The three new points of light near Jupiter) If...the little points of light near Jupiter are caused by immovable stars embedded in the external celestial sphere, (and)...three new points of light were similarly seen in the night sky, (then) perhaps they also are fixed stars. Invention of hypothesis < Fixed stars.>

Retroduction: Does the proposed explanation extend to what we already know? If...the points of light are fixed stars, (and)...their positions are compared to each other, (then) their positions should be random. (But)...they appear exactly in a straight line, parallel to the ecliptic. (Therefore) perhaps as Galileo put it, "yet they made me somewhat wonder".

Selection of hypothesis <suspected Fixed stars>

> Although I believed them to belong to the number of the fixed stars, yet they made me somewhat wonder, because they seemed to be arranged exactly in a straight line, parallel to the ecliptic, and to be brighter than the rest of the stars. The position of them with reference to one another and to Jupiter was as follows (Originally in [3], p. 59).
> (east) $*, *, *, o, *$ (west)

Yet they made him somewhat wonder. Thus, the need to test his hypothesis again became apparent. Rather than reject the wonder hypothesis, he decided to test the hypothesis again for more data, or test's faults at that time.

The testing of hypotheses

> This comparison allows one to draw a conclusion. A good match means that the hypothesis is supported, but not proven. While a poor match means that something is wrong with the hypothesis, the test, or with both. In the case of a good match, the hypothesis has not been "proven" correct with certainty because one or more unstated and perhaps un-imagined alternative hypotheses may give rise to the same prediction under this test condition (e.g., [5, 31]. Similarly, a poor match cannot "disprove" or falsify a hypothesis in any ultimate sense. A poor match cannot be said to falsify with certainty because the failure to achieve a good match may be the fault of the test condition(s) rather than the fault of the hypothesis (Originally in [5, 31].

Deduction—Induction: Hypothetico-deduction: The next evening (January 8) Galileo turned his telescope on the planet again, hoping to see that it had moved to the west of the stars, as astronomical computations then predicted. To his surprise, this time he found it to the east of them.

If...the points of light are fixed stars, (and) Jupiter was then moving in a westwards ("retrograde") motion, and he observed them over the next several nights.

(Then) Galileo expected that the following night Jupiter would be west of the supposed stars.

(But) in fact, it was to the east. (Therefore) they are not fixed stars.

> When Galileo turned his newly adapted spyglass to it on the evening of 7 January, his attention was drawn to the formation shown earlier. He thought, of course, he was seeing three little fixed stars in a row and that Jupiter just happened to be passing through their formation that evening. Near opposition, Jupiter's motion with respect to the fixed stars is

> retrograde, that is, from the east to the west, and therefore, when Galileo again sought out Jupiter, he expected to see the stars in the same formation, with Jupiter having moved to the west with respect to them. What he saw in fact was that Jupiter had moved to the east, still on the same straight line. This puzzled him, and he thought that perhaps the astronomical tables were wrong and Jupiter had returned to its direct, west to east, motion. Jupiter's seemingly anomalous behavior greatly intrigued him. They answered a major criticism against the Copernican theory: if the Earth were a planet, why should it be the only one to have a moon going around it, and how could there be two centers of motion in the universe? (Originally in Galileo 1564–1642, pp. 15–16).

4.2 Suggested and Selected Moon Hypothesis

From effect to cause; the cause, is confirmed, eliminating other possibilities

Effect: Four little stars accompany Jupiter, always in a straight line with it, and move along the line with respect to each other and to Jupiter.

Cause: Moon hypothesis is materially recognized formally as the cause of the effect.

Recycling: Analogical abduction: non-monotonic cycle (Route 2)

The generation of hypotheses

Abstraction from observed data by theory-dependent method: A number of observations were made between January 7 and 15, analyzing in detail their variation in position, how they are separated from Jupiter or each other and merged with them in successive observations; inference to the only possible motion that explains theses details; concluding "no one can doubt" (nemini dubium esse potes) that they complete revolutions around Jupiter in the plane of the elliptic, each at a fixed radius and with its characteristic time of revolution (GG3.1 94).

Analogical abduction: If...the points of light near a planet in the night sky are caused by moons of a planet moving in the celestial sphere, (and)...three such points of light are observed near Jupiter in the night sky, (then)...perhaps they also are the moons orbiting Jupiter. *Invention of hypothesis <The moons orbiting Jupiter>*

Retroduction: Application of Lakatos' criterion requires historical inquiry. The philosopher of science must survey the science to see whether there exist alternative hypotheses that do not imply the evidence. Old evidence provides support only within the context of competition between hypotheses [15, p. 227]. The Selection of a hypothesis <The moons orbiting Jupiter hypothesis > , eliminating < fixed stars hypothesis >

If...the three points are moons orbiting Jupiter,
(and)...he observed them over the next several nights,
(then) they should appear along a straight line on either side of Jupiter.
(And) they appear exactly in the straight line parallel to the ecliptic on either side of Jupiter.
(Therefore) perhaps they are moons orbiting Jupiter. <The moons orbiting Jupiter>

If...the points of light are fixed stars,
(and)...their positions are compared to each other,
(then) their positions should be random, or they can appear in the straight line parallel by chance.
(But)...they appear in the straight line approximately.
(Therefore) perhaps the fixed stars hypothesis is not supported. <Fixed stars>

Updating Hypothesis: Perhaps they are moons orbiting Jupiter. Some objects within the celestial sphere—the planets—orbit the sun (e.g., Earth, Venus, Jupiter). Galileo's knowledge about heavenly objects would have included this assumption [13]. Thus, the stars must in fact be satellites, moons circling around Jupiter and carried along by the planet, just as the moon orbits Earth.

The testing of hypotheses
Deduction—Induction: Hypothetico-deduction

What does the proposed explanation lead us to predict about further observation?
If...these points of light are moons orbiting Jupiter,
(and)...I observed them over the next several lights,
(then) some nights they should appear to the east of Jupiter and some nights they should appear to the west. Further, they should appear along a straight line on either side of Jupiter. Deduction

How are the predictions and new observations compared?

(And)...the new observations match the predictions based on the orbiting-moons hypothesis (e.g., some nights they appeared to the east of Jupiter and some nights they appeared to the west),

therefore...the orbiting-moons hypothesis is supported. Induction

Galileo's most striking discovery concerned the planet Jupiter [6]:

> When he first examined it on January 7, 1610, he found the planet in the midst of three little stars ranged—curiously—in a straight line. Jupiter was then moving in a westwards ("retrograde") motion, and Galileo therefore expected that the following night (January 8) Jupiter would be the west of the supposed stars; but in fact it was to the east. The next night was cloudy, but on January 10, he found the planet to the west of two stars, with the third star nowhere to be seen. By January 13, the number of stars had increased to four; and by January 15, Galileo had realized that the supposed stars must in fact be satellites, moons circling around Jupiter and carried along by the planet as it orbited the Sun [6, p. 125].

4.3 *Moon hypothesis Expansion from the Cause, Recognized Formally as the Cause, to its Proper Effects*

Cause: Four satellites of Jupiter always accompany Jupiter, in direct and retrograde motion, with their own distances from it and periods of revolution, as it revolves around the center in about 12 years.

Effect: At certain edge the appearance of four points of light, moving back and forth on a line with the planet(Jupiter) and parallel to the elliptic.

Recycling for expansion: Updating—Hypothetico deduction Cycle (Route A) It employs similar suppositions, mainly taken from projective geometry and geometrical optics, and of course is unintelligible to those unacquainted with those disciplines. <*Updating*>

More importantly, it supposes that the observational evidence presented by Galileo is correct and that it can be telescopic. This supposition definitely showed the acceptance of the verification and is accepted by the scientific community in a short time. On March 24, 1611, Jesuit astronomers at the Collegio Romano confirmed Galileo's discovery <*Hypothetico deduction Cycle*>, writing to their confrere, Cardinal Bellarmine, that [35, p. 203]:

> Four stars go about Jupiter, which move very swiftly, now all to the east, and now all to the west, and sometimes some move to the east and some to the west, all in an almost straight line. These cannot be fixed stars, for they have very swift motions, very different from those of the fixed stars, and they always change their distances from each other and from Jupiter (GG 11. 93).

Even in the Ptolemaic universe, Jupiter moved, and the telescope showed that it had moons and kept them. Ptolemaic astronomers argued that Earth could not move because it would lose its moon. However, evidently, Earth did not leave its moon behind. Therefore, Earth can move, too.

Galileo presented his arguments in the form of evidence and conditions, and the moons of Jupiter were key evidence. Ptolemaic astronomers argued that Earth cannot move or it would lose its moon, but even in the Ptolemaic universe, Jupiter moved, and the telescope showed that it had moons and kept them. Evidently, Earth could move and not leave its moon behind. Furthermore, moons circling Jupiter did not fit the classical belief that all motion was centered on Earth. Obviously, there could be other centers of motion. Finally, the orbital periods of the moon were related to their distance from Jupiter, just as the orbital periods of the planets were, in the Copernican system, related to their distance from the sun. This similarity suggested that the sun rule its harmonious family of moons.

The satellites of Jupiter by Galileo were another matter, perhaps [36]:

> Before their discovery, the moon, a planet circling a planet as it were, had appeared to be an unexplained anomaly in the heliocentric system and therefore an objection to it. If the satellites of Jupiter did not explain the phenomenon, at least they destroyed its uniqueness, and the moon appeared to be less anomalous. The satellites of Jupiter offered no positive support for the heliocentric system, however. The phases of Venus did. There was one other expected phenomenon that the telescope did not reveal, however, and as far as the Copernican revolution is concerned, it was the most perplexing telescopic observation..The telescope did not reveal stellar parallax. From the moment when the Copernican system was born, the crucial relevance of stellar parallax had been obvious. Galileo's telescope could not distinguish it, and non-appearance of stellar parallax balanced, at the very least, the positive evidence offered by the phase of Venus. The case for the Copernican-Keplerian system stood or fell on the argument of geometric harmony and simplicity. For the advantage and for little else, men were asked to overturn a conception of the universe that included physical, philosophical, psychological, and religious questions

> of the most all-embracing nature. Perhaps, it was more of a load than geometric simplicity could bear (Originally in [36, pp. 13–15]).

(If...) Earth could move, too, (and...) I make observations over the next several years through the telescope, (then) parallax requires that in summer they should appear to the radial angle of a star, and in winter they should appear to the radial angle of the same star.

(But) the new observations did not match the predictions of heliocentric hypotheses based on the hypothesis.

(Therefore) the heliocentric hypothesis was not supported at that time.

> Not the least of the sacrifice demanded in the name of simplicity was common sense itself. It has been remarked many times that modern science has required a re-education of common sense. What could have been more common-sensical than a geocentric universe? We still say that the sun rises and speak of a solid earth. The heliocentric universe demanded that plain evidence of the sense in such matters be denied as mere illusion (Originally in [36, pp. 15–16].

Many people chose heliocentric hypothesis due to its mathematical simplicity. But this simplicity as background knowledge didn't give help enough to support main hypothesis(H, Galileo's Heliocentric Hypothesis) Because at that time through the low magnified telescope, the time difference(the auxiliary hypothesis A) were not be discovered. Then, we could conclude that Galileo's heliocentric universe was false, but A was correct. Thus, it turned out that Tycho-Model based on (A) geocentric universe involving those observational data was supported better.

If new information emerges *(Route B)* during the previous cycle, then a new Abduction-Deduction-Induction cycle must be undertaken to achieve more plausible (radical dynamic and static consilience) hypotheses, after this Recycle.

5 Conclusions and Discussions

In this study, we have proposed a noble and refined model for abductive inferences inspired by Peirce. And we have shown its validity by applying to Galileo's discovery of the moons of Jupiter in history of science.

First, we examined abductive inference proposed by Charles Sandra Peirce. We suggested a scientific inference procedure based on abductive inference strategies that could be used to generate hypotheses. These abductive inferences are not static but dynamic. They involve cycles and recycles. The term "abduction" is usually applied to the evaluation of explanatory hypotheses, but it sometimes also refers to the process of generating hypotheses. Because the hypotheses are invented via analogical abduction to explain puzzling phenomena are only tentative hypotheses, the role of creativity is paramount.

Second, to select most plausible hypothesis or theory, we need evaluation(called weak evaluation in this article). The first step for evaluation is

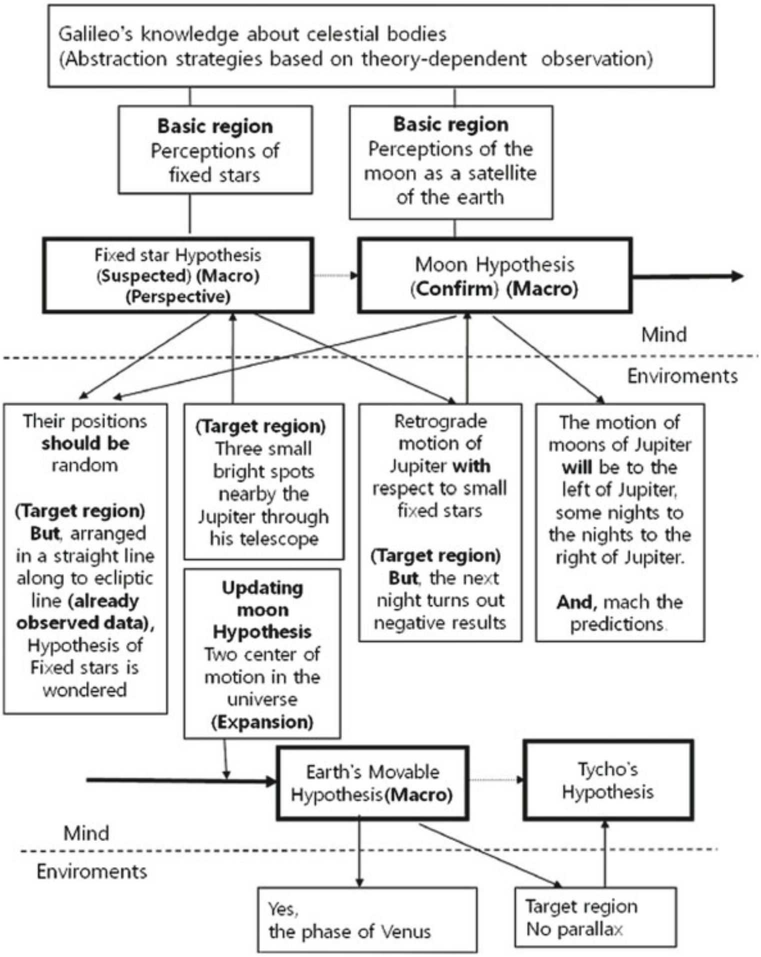

Galileo's knowledge about celestial bodies
(Abstraction strategies based on theory-dependent observation)
Basic region
Perceptions of fixed stars
Basic region
Perceptions of the moon as a satellite of the earth
Fixed star Hypothesis (Suspected) (Macro) (Perspective)
Moon Hypothesis (Confirm) (Macro)
Mind
Enviroments
Their positions should be random
(Target region) But, arranged in a straight line along to ecliptic line (already observed data), Hypothesis of Fixed stars is wondered
(Target region) Three small bright spots nearby the Jupiter through his telescope
Retrograde motion of Jupiter with respect to small fixed stars
(Target region) But, the next night turns out negative results
The motion of moons of Jupiter will be to the left of Jupiter, some nights to the nights to the right of Jupiter.
And, mach the predictions.
Updating moon Hypothesis Two center of motion in the universe (Expansion)
Earth's Movable Hypothesis(Macro)
Tycho's Hypothesis
Mind
Enviroments
Yes, the phase of Venus
Target region No parallax

Fig. 2 Representation of Galileo's discovery of the moons of Jupiter. The *upward arrows* represent analogical abductive inference, whereas the *right downward arrows* represent retroductive and *left downward arrows* represent predictive inference, and the *arrows* to the *right* represent the updating stage including new data or theories

retroduction and updating. Updating involves new available data to develop more plausible hypotheses. Second step for evaluation and selection, we have suggested the hypothetico-deduction, and induction (deduction-Induction). This deduction-induction processes are practical procedures for evaluating and selecting hypotheses. In Fig. 2, these procedures are described clearly.

Third, we have applied our model to Galileo's discovery of moon orbiting Jupiter. The data Galilleo acquired through observation were used to selection of hypothesis. The selected hypothesis is old hypothesis, i.e. fixed-star hypotheses, and it was suspected.

By using Retroduction and Hypothetico-Deduction(Deduction-Induction), the fixed-star hypotheses were rejected by Galilleo. And new generated hypotheses are Moon hypotheses through analogical abduction, retroduction and updating. This is the procedure from effect to cause. The effect is "four little stars accompany Jupiter, always in a straight line with it, and move along the line with respect to each other and Jupiter". The cause for the effect is Moon hypothesis [35, p. 203].

The starting point of the inquiry occurs when "*the cause of the effect is materially suspected*". In the progression from effect to cause, the cause is materially suspected but is not yet recognized formally as the cause. The intermediate stage involves the intellectual work by eliminating other possibilities to determine whether this is the real cause of the effect. We propose that both the "from effect to cause" stage and the "intermediate stage" involve both "*the generation of hypotheses*" and "*the testing of the hypotheses*". It makes three broad macro perspectives; (1) "fixed stars hypothesis suspected", (2) Moon hypothesis can be suggested and selected, (3) Moon hypothesis expansion.

Natural scientists have used these abductive inference patterns in various theories for causal explanations for natural phenomena based on abstraction strategies through theory-laden observation. These include [10, 12]. We have successfully shown that our noble model for abduction inferences is very good in accounting for Galilleo's discovery of Moons orbiting Jupiter. Our further research is to refine this model and prove its validity through applying to other historical cases.

References

1. Aliseda, A.: Abductive Reasoning: Logical Investigations into Discovery and Explanation. Springer, Dordrecht (2006)
2. Archer, R., Bhaskar, R., Collier, A., Lawson, T., Norrie, A.: Critical Realism: Essential Readings. In: Archer, M., Bhaskar, R., Collier, A., Lawson, T., Norrie, A. (eds.) Routledge, London (1998)
3. Galilei, G.: The sidereal messenger. In: Shapley, H., Rapport, S., Wright, H. (eds.) (1954), A Treasury of Science. Harper & Brothers, New York (1610)
4. Hanson, N.R.: Patterns of Discovery. Cambridge University Press, London (1972)
5. Hempel, C.: Philosophy of Natural Science. Prentice-Hall, Upper Saddle River (1966)
6. Hoskin, M.: The Cambridge Illustrated History of Astronomy. Cambridge University Press, Cambridge (1997)
7. Josephson, J.R., Josephson, S.G.: Abductive Inference: Computation, Philosophy, Technology. Cambridge University Press, Cambridge (1996)
8. Kapitan, T.: Peirce and the autonomy of abductive reasoning. Erkenntnis **37**, pp. 126 (1992)
9. Kapitan, T.: Peirce and structure of abductive inference. In: Houser, N., Roberts, D.D., Evra, J.V. (eds.) Studies in the Logic of Charles Sanders Peirce, pp. 477–496. Indiana University Press, Bloomington and Indianapolis (1997)
10. Kuhn, T.S.: The Structure of Scientific Revolutions, 2nd edn. University of Chicago Press, Chicago (1970)
11. Lakatos, I.: Changes in the problem of inductive logic. In: I. Lakatos (ed.) Inductive Logic, pp. 376–377. North-Holland Publishing, Amsterdam (1968)

12. Lakatos, I.: Falsification and the methodology of scientific research programmes. In: Lakatos, I., Musgrave, A. (eds.) Criticism and the Growth of Knowledge. Cambridge University Press, New York (1970)
13. Lawson, A.: What does Galileos discovery of Jupiter's moons tell us about the process of scientific discovery? Sci. Educ. **11**, 1–24 (2002)
14. Lawson, A.E.: Basic Inferences of Scientific Reasoning, Argumentation, and Discovery. Sci. Educ. **94**, 336–364 (2010)
15. Losee, J.: A Historical Introduction to the Philosophy of Science, 4th edn. Oxford University Press Inc, New York (2001)
16. 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)
17. Magnani, L.: Epistémolgie de I ' invention scientifique. Communication and Cognition. **21**, 273–291 (1988)
18. Magnani, L.: Abductive reasoning: Philosophical and educational perspectives in medicine. In: Evans D.A., Patel V.L. (eds.) Advanced Models of Cognition in Medical Training and Practice, pp. 21–41. Springer, Berlin (1992)
19. Magnani, L.: Abduction, Reason, and Science Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
20. McLaughlin, R.: Invention and induction: Laudan Simon, and the logic of discovery. Philos. Sci. **49**, 209 (1982)
21. Niiniluoto, I.: Defending abduction. Philos. Sci. (proceedings) **66**, S436–S451 (1999)
22. Oh, J.-Y.: Defending problems with Peirce's concept of abduction. J. Korean Philos. Soc. **113**, 215–255 (2010)
23. Oh, J.-Y.: Understanding scientific inference in the natural sciences based on abductive inference strategies. In: Magnani L., Li P. (eds.) Philosophy and cognitive science: Western & Eastern studies (Sapere 2), pp. 221–237. Springer, New York (2012)
24. Paavola, S.: Abduction as a logic and methodology of discovery: the importance of strategies. Found. Sci. **9**, 267–283 (2004)
25. Paavola, S.: Peircean abduction: Instinct or inference? Semiotica **153**(1/4), 131–154 (2005)
26. Peng, Y., Reggia, J.A.: Abductive Inference Models for Diagnostic Problem-Solving. Springer, New York (1990)
27. Pera, M.: Inductive method and scientific discovery. In: Grmck, M.D., Cohen, R.S., Cimino, G. (eds.) On Scientific Discovery. D.Reidal Pub. Co, Dordrecht (1981)
28. Ramoni, M., Stefannelli, M., Magnani, L., Barosi, G.: An epistemological framework for medical knowledge-base systems. IEEE Trans. Syst. Man Cybern. **22**(6), 1361–1375 (1992)
29. Rescher, N.: Peirces Philosophy of Science. University of Notre Dame Press, Indiana (1978)
30. Salmon, W.C.: The foundations of scientific inference. University of Pittsburgh Press, Pittsburgh (1967)
31. Salmon, M.: Introduction to logic and critical thinking (3rd ed.). Harcourt Brace, Fort Worth, TX (1995)
32. Thagard, P.: The best explanation: criteria for theory choice. J. Philos. **75**, 76–92 (1978)
33. Thagard, P.: Computational Philosophy of Science. MIT Press, Cambridge (1988)
34. Trigg, G.I. (ed.): Encyclopedia of Applied Physics, vol. 2. VCH Publishers Inc, New York (1991)
35. Wallace, W.A.: Galileos Logic of Discovery and Proof: The Background, Content, and Use of His Appropriated Treatises on Aristotles Posterior Analytics. Kluwer Academic Publishers, Dordrecht (1992)
36. Westfall, R.S.: The Construction of Modern Science. Wiley, New York (1971)
37. Whewell, W.: The Philosophy of the Inductive Sciences. John W. Parker, London (1847)
38. White, M.: Galileo Antichrist: A Biography. Weidenfeld & Nicolson, London (2007)
39. Wuisman, J.J.J.M.: The logic of scientific discovery in critical realist social scientific research. J. Crit. Realism **4**(2), 366–394 (2005)

Author Biographies

Jun-Young Oh is assistant professor at the Hanyang University, South Korea. He received his BA in Science Education from Kongju National Teachers’ College, South Korea, his MS in Astronomy from Yonsei University, South Korea, and his PhD in the Philosophy of Science & Science Education from Pusan National University, South Korea. His research focuses on Cognitive Science based on Abduction and Science Education based on the Philosophy of Science.

Yoo Shin Kim is full Professor, Chairman of Graduate interdisciplinary program Science & Technology Studies at Pusan National University, South Korea. He received his BA in Electronics Engineering from Seoul National University, South Korea, his MS in Dept. of EECS from U.C. Berkeley, USA, and his PhD in the Philosophy of science from Cornell University, USA. His research focuses on Realism–Anti realism debate, Neural Network, Philosophy of Quantum Mechanics, and Philosophy of Culture.

Icon and Abduction: Situatedness in Peircean Cognitive Semiotics

Pedro Atã and João Queiroz

Abstract Differently from the anti-cartesianism defended by some embodied-situated cognitive scientists, which is predominantly anti-representationalist, for C. S. Peirce, mind is semiosis (sign-action) in a dialogical form, and cognition is the development of available semiotic material artifacts in which it is embodied as a power to produce interpretants (sign-effects). It takes the form of development of semiotic artifacts, such as writing tools, instruments of observation, notational systems, languages, and so forth. Our objective in this paper is to explore the connection between a semiotic theory of mind and the conception of situatedness and extended mind through the notions of iconicity and abductive inference, taking advantage of an empirical example of investigation in distributed problem solving (Tower of Hanoi).

1 Introduction

Charles S. Peirce can be considered an important precursor of situated mind and distributed cognition thesis. But differently from the anti-cartesianism defended by some embodied-situated cognitive scientists, which is predominantly anti-representationalist, as recently explored in a Merleau-Pontyan [1], Heidegerian [2], or a Gibsonian [3] trend, for Peirce, mind is semiosis (sign-action) in a dialogical—hence communicational—materially embodied form, and cognition is the development of available semiotic material artifacts in which it is embodied as a power to produce interpretants. It takes the form of development of semiotic artifacts, such as writing tools, instruments of observation, notational

P. Atã · J. Queiroz (✉)
Institute of Arts and Design, Federal University of Juiz de Fora, Juiz de Fora, Brazil
e-mail: queirozj@pq.cnpq.br

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_17, © Springer-Verlag Berlin Heidelberg 2014

systems, languages, and so forth, as stressed by Skagestad [4] with respect to the concept of intelligence augmentation.

Although only recently a more systematic discussion upon the distributed nature of the mental processes have been established in empirical fields (e.g. neurocognitive science, artificial intelligence), the philosophical basis of this thesis and its variations have well-known predecessors. Among them, the most quoted are William James, Wittgenstein, John Dewey, James Gibson, Vigotsky, Merleau-Ponty, Heidegger (see [2, 5]). However, Charles Sanders Peirce, the least mentioned among the pragmatists in this context, can be considered an avant-garde situated and embodied cognition proposer. In Peircean Semiotic Theory of Mind the fundamental unit of cognitive interest is reconceived—disembodied mind is replaced by environmentally embedded space of semiotic skills and artifacts.

Our objective in this work is to explore the connection between a semiotic theory of mind and the conception of situatedness through the notions of iconicity and abductive inference, taking advantage of an empirical example of investigation in distributed problem solving (Tower of Hanoi). In the following sections we introduce: (i) the basic semiotic relations that ground a semiotic theory of mind, (ii) the notions of iconicity and abductive inference as specially near to the conceptualization of situatedness and distributedness of reasoning, (iii) the experiment of the Tower of Hanoi, conducted by Zhang and Norman [6], analyzed through the framework provided.

2 Semiosis and Semiotic Theory of Mind

Peircean approach of semiotic processes (semiosis) is related to formal attempts to describe cognitive processes in general. This framework provides: (i) a pragmatic model of semiosis, (ii) a conception of mind as a sign-interpretation process (see [7]), and (iii) a list of fundamental varieties of representations based on a theory of logical-phenomenological categories.

According to the Peircean model, a meaning process involves a relational complex constituted by three terms *irreducibly* connected—Sign, Object and Interpretant (S–O–I). The *irreducibility* indicates a logical property of this complex: the sign process must be regarded as associated to the *interpretant*, as an ongoing process of interpretation [8], and is not decomposable into any simpler relation (CP 5.484). Peirce also defines a sign as a medium for the communication of a *form* or *habit* embodied in the object [9, 10]. This *form* is communicated to the interpretant, so as to constrain (in general) the interpretant as a sign or (in biological systems) the interpreter's behavior. The object of sign transmission is a habit (a regularity, a rule of action, or a 'pattern of constraints') embodied as a constraining factor of interpretative behavior—a logically 'would be' fact of response. The habit embodied in the object allows a semiotic system to interpret the sign as indicative of a class of entities or phenomena [11]. Meaning and meaning change are conceived as a constraining factor of possible patterns of

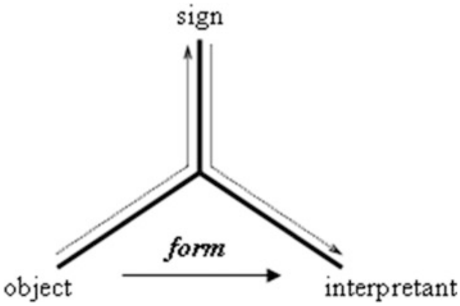


Fig. 1 Semiosis as a relation between three irreducibly connected terms (sign-object-interpretant, S–O–I). This triadic relationship communicates a form from the object to the interpretant through the sign (symbolized by the *horizontal arrow*). The other *two arrows* indicate that the form is conveyed from the object to the interpretant through a determination of the sign by the object, and a determination of the interpretant by the sign

interpretative behavior through habit and change of habit. The mediation of a sign results in a consistent relationship between variations in the form of the object and the corresponding effects on the interpreter (Fig. 1).

Sign-mediated processes show a remarkable variety. The construction of appropriate typologies of these processes is a requisite for a deeper and more refined understanding of cognition. In an attempt to advance in the understanding of semiotic processes, Peirce proposed several typologies, with different degrees of refinement and several relationships to one another. A basic typology in his framework differentiates between iconic, indexical, and symbolic processes.

3 Fundamental Kinds of Signs: Icons, Indices, and Symbols

Icons, indices, and symbols are differentiated by Peirce based on how the sign relates to its object, that might be defined as the item to which the interpretants are related by the mediation of sign (see [12]). This typology exhibits a property capable of functioning as an operational criterion to distinguish different kinds of signs: the relative dependence of sign-object-interpretant (S–O–I) components in triadic relation [13, 14].

A symbol is an S–O relationship logically dependent of I. This relation has been characterized as a law ascribing S–O. A symbol is “a Sign (q.v.) which is constituted a sign merely or mainly by the fact that it is used and understood as such, whether the habit is natural or conventional, and without regard to the motives which originally governed its selection” (CP 2.307). Differently, an index is dependent of O. The relation between S and O has been characterized as one of contiguity: constraints resulting from the space–time existence of the object—irrelevant in symbolic processes—are the reason for the representation of O

Table 1 The fundamental types of signs underlying meaning processes—icons, indexes, and symbols

Type of sign	S–O relation	S–O–I dependence	
Icon	Similarity	Monadic (S)	Dependent of intrinsic properties of S
Index	Contiguity	Dyadic (S–O)	Dependent of S–O spatio-temporal correlation
Symbol	Law	Triadic (S–O–I)	S–O dependent of I mediation

They are characterized in terms of relative dependence of sign-object-interpretant (S–O–I) components in triadic relation. The icon is the sign whose relevant properties for signification are its own intrinsic qualities: S depends on S

through S. In that case, S is really determined by O, in such a way that both must exist as events. The notion of spatio-temporal co-variation is the most characteristic property of indexical processes. When S is an icon, S signifies by means of qualities of S. Icons are dependent on the material, form and structure that are made–"An Icon is a sign which refers to the Object that it denotes merely by virtue of characters of its own, and which it possesses, just the same, whether any such Object actually exists or not" (CP 2.247). This relation between S and O based on the qualities of S has been characterized as one of similarity. The problem with the notion of similarity, however, is that it is too vague (see [15]). In order to detrivialize the notion of icon as a sign based on similarity it is possible to give an operational definition of the icon (Table 1).

4 Iconicity: Operational Notion

The icons' dependence of its own materiality makes them suitable for modeling and experimentation. When an *operational criterion* is adopted, the icon is defined as anything whose manipulation can reveal more information about its object. Algebra, syntax, graphs, and formalizations of all types should be recognized as icons. This definition is considered a detrivilization of the notion that the icon is fundamentally based on a relation of similarity (see [15]; also [16]).

> The key of iconicity is not perceived resemblance between the sign and what it signifies but rather the possibility of making new discoveries about the object of a sign through observing features of the sign itself. Thus a mathematical model of a physical system is an iconic representation because its use provides new information about the physical system. This is the distinctive feature and value of iconic representation: a sign *resembles* its object if, and only if, study of the sign can yield new information about the object [16, p. 102].

The icon is notably related to situatedness and distributedness of reasoning. It is the sign whose signification is S-dependent (that means, dependent on the sign itself) and allows, through its manipulation, some discovery about the object. The notion of iconicity attests the capacity of material features to be the semiotic basis of cognitive operation, and not only play a secondary role.

5 Abduction: First Stage of Inquiry

Inferences are also understood as semiotic processes and have a place reserved under Peirce's typology. They are classified into three irreducible types –abduction, deduction and induction—corresponding to three subsequent phases in the process of scientific inquiry (CP 6.469-473). Abduction rises from the observation of a mass of facts that doesn't fit into the habits and expectations of the observer and culminates with the formation and selection of a hypothesis. Deduction develops testable consequences of the previously generated hypothesis. Based on these consequences, induction performs tests to evaluate it.

The characterization of abduction as the transformation of mass of facts into hypotheses and the first stage of inquiry brings it close to perception (see [17, 18]). For Peirce, perception involves an interpretative process (CP 5.181). It is through an inferential-like perceptual judgment that percepts are subsumed under general classes. This perceptual judgment accounts for the transformation of sense data into knowledge applicable to theoretical or practical use. It is subconscious, but if it was subjected to logical analysis, it would present an inferential—abductive—form (CP 5.181). Therefore, "all that makes knowledge applicable comes to us via abduction" (MS 692).

As an "act of insight" that "comes to us like a flash" (CP 5.181) abduction is germane to creativity. For Peirce, abduction is also the logical inference by which new knowledge can be obtained: "Abduction consists in studying the facts and devising a theory to explain them. It is the only kind of argument which starts a new idea" (CP 2.96). Magnani [19] introduces the concept of "manipulative abduction" to refer to those cases where the inference depends on the exploration of external resources—it "happens when we are thinking *through* doing and not only, in a pragmatic sense, about doing" [19, p. 274]. According to Paavola [20], in abduction the iconic character of reasoning is more prominent. Icons, abductions and perceptual judgments all have important similarities between themselves.

> In all of them, some characteristics or phenomena suggest a potential way of interpreting or explaining these characteristics or phenomena and bringing them into some kind of an order [20, p. 305]

Paavola has referred to these characteristics that only *suggest* a way in which they could be interpreted as *clue-like* characteristics. In abduction, these clue-like characteristics, together with background knowledge, lead to the conclusion of a hypothesis (i.e., a promising way of arranging a mass of facts). This is a distributed process whenever these *clue-like characteristics* are predominantly material qualities of external signs. Abduction is especially near to the conceptualization of distributedness because it is an inference which relies on a mass of percepted data for its conclusion.

To see how iconicity and abduction are related to situatedness, we analyze in the next section an example of distributed reasoning. More specifically, we identify the role of both icons and abductions in the distributed problem solving task of the Tower of Hanoi.

6 Externalization of Constraints as an Iconic-Embedded Abductive Process

The Tower of Hanoi is a puzzle game. It is (normally) constituted of three poles and several disks of variable diameters with a hole in the centre in order to be stacked in the poles (see Fig. 2). The diameter of the disks represents the hierarchy according to which they may be organized or moved across the poles. The goal of the game is to rearrange the disks from a specific initial state to a specific goal state, while observing some basic rules. The formal structure of the game is composed by the pieces (disks, for example), places (poles), hierarchy (disk diameters), rules, initial state, and goal state.

Zhang and Norman [6] have used the tower of Hanoi game to study the influence of representations in cognition. More specifically, they were dealing with the Representational Effect: difference in cognitive behavior caused solely by representational features. The Representational Effect is investigated through the comparison of performance upon isomorphic representations in problem solving tasks, i.e., representations that carry the same amount of information, but that vary in the way that this information is presented. In the experiment treated here, the authors have used the isomorphic versions of the Tower of Hanoi puzzle showed in Fig. 3.

Zhang and Norman's tests covered several levels of isomorphism between representations (level of object representations, level of dimensional representations, level of rule representations and level of problem space structures). The particular experiment that interests us (experiment 2, Zhang and Norman [6],

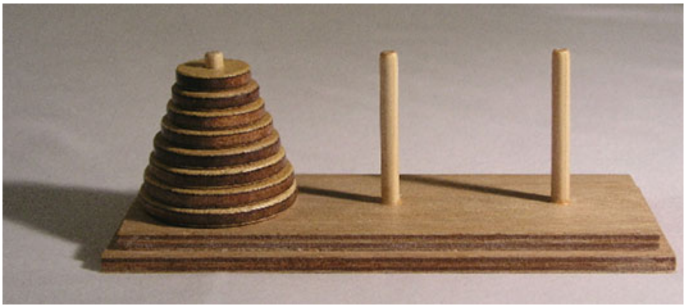


Fig. 2 The classical version of the Tower of Hanoi puzzle, with three poles and several disks stacked from the largest, in the base, to the smallest, in the top. In the experiments treated here, this order was altered: larger pieces should be put on top of smaller pieces. Image taken from Wikimedia Commons

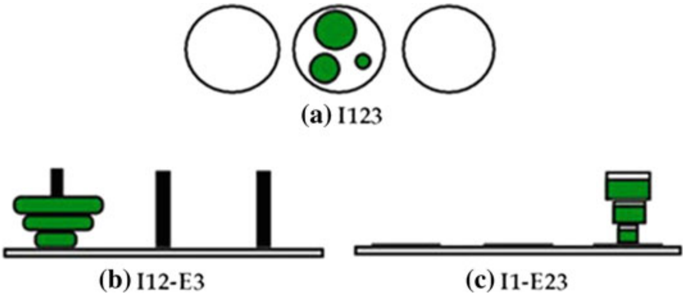


Fig. 3 Three isomorphs of the tower of Hanoi which vary in respect to the externalization of constraints. In **a** the three rules of the game are internal. In **b** two of the rules are internal and one is external. In **c** only one of the rules is internal, and the other two are external [6]

pp. 20–23) is the level of rule representations. In this level, the rules of the game itself can be represented in two ways: they are either (i) stated in instructions and memorized by the players or (ii) automatically embedded in the possibilities of move offered by the material of play. Rules introduced according to (i) and (ii) are termed, respectively, *internal* and *external* rules, kept, in the act of play, either in the memory of the players or in the material of play itself.

There were three rules in the game for this experiment (see Table 2) and two ways in which these rules could be introduced (internal or external rules). Three isomorphs were used (see Table 3) the, "waitresses and oranges", "waitresses and donuts" and "waitresses and coffee", that differently represent the elements that compose the formal structure of the game. The oranges version utilizes balls ("oranges") as the pieces, plates as the places and the size of the balls as the hierarchy. The donuts version utilizes disks ("donuts") as the pieces, poles as the places and the diameter of the disks as the hierarchy. The "coffee" version utilizes cups filled with coffee as the pieces, plates as the places and the size of the cups as the hierarchy. Each of the three rules were either internal (given as a list of

Table 2 Rules of the TOH, experiment 2

1. Only one piece can be transferred at a time
2. A piece can only be transferred to a place on which it will be the largest
3. Only the largest piece in a place can be transferred to another place

Table 3 Isomorphic representations of the game's formal structure

	"Oranges" (I123)	"Donuts" (I12 E3)	"Coffee" (I1 E23)
Pieces	Balls	Disks	Cups filled with coffee
Places	Plates	Poles	Plates
Hierarchy	Size of balls	Diameter of disks	Sizes of cups
Rules	1. Instruction	1. Instruction	1. Instruction
	2. Instruction	2. Instruction	2. Material
	3. Instruction	3. Material	3. Material

instruction read before the experiment and memorized by the players) or external (automatically embedded in the material of play). In the "oranges" version, all the three rules were internal (I123). In the "donuts" version, rules 1 and 2 were internal and rule 3 was external (I12 E3). In the "coffee" version, only rule 1 was internal and rules 2 and 3 were external (I1 E23). The oranges version is internal in respect to all rules because the balls in plates can be physically moved without any constraining in relation to each other. The donuts version is external in respect to rule 3 because the stacking of disks in poles only allow that the disk in top be physically moved (unless you take more than one disk, but in this case you would be breaking the internal rule 1). The coffee version is external in respect to rules 2, 3 because, beyond being stacked one on top of the other (rule 3), a smaller cup, filled with coffee, cannot be placed on top of a bigger cup, filled with coffee, because in this case the coffee will spill. In a context where it is understood that spilling coffee is bad, rule 2 has also been externalized.

The experiment measured the time required for solution, the number of steps required for solution and the number of wrong moves for each of the three isomorphs. In the three cases, the results for the most internalized version (oranges) were the worst: more time to solve, more number of steps required to solve and more wrong moves. For the most externalized (coffee) the results were the best: less time to solve, less number of steps required and almost no wrong moves. The donuts version stayed in the middle (see Fig. 4). This experiment, together with others in the same article, have led the authors to propose that more externalized representations are also more efficient representations for problem solving (see also [21, 22]).

The criterion the authors have used to classify between internal and external rules matches a criterion for iconicity, namely, dependence of material properties, i.e. S-dependence. The different isormophs of the experiment can be modeled as semiotic processes of communication of a form or habit from an object to an interpretant through the mediation of the sign. The object (O) of this triadic relation is the formal structure of the game that is common to all isomorphs. The sign (S) is the medium through which the game is played, i.e., the specific pieces and places and also the list of written instructions. The interpretant (I) is the constraining in behavior that characterizes the act of play itself. With this

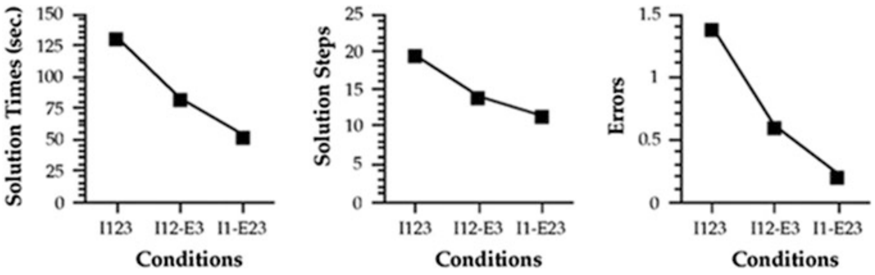


Fig. 4 Results of the experiment for each of the isomorphs [6]

framework in mind, and taking into consideration the criterion of relative dependence of terms for the fundamental classification of signs, we conclude that, for the (i) internal and (ii) external cases:

(i) O (formal structure of the game) is independent of S (material of play). If you change the materials used to play, the game remains the same. The S–O relation cannot be established by these two terms alone, it requires the mediation of a third term (I). The constraining upon the specific material of play, that makes it correspond to the formal structure of the game, only happen as a cognitive constraining in the behavior of the player, in the act of play itself. As S–O relation is dependent of I, this is an example of symbolic semiosis.
(ii) The game is S-dependent. If you change the materials used to play, the formal structure of the game changes. The S–O relation is already established independently of the third term (I), because the constraints of S are a materialization of the formal structure of the game. The constraining upon the specific material of play, that makes it correspond to the formal structure of the game, is already given in the material of play, before the game is played. As S–O is dependent of S, this is an example of iconic semiosis.

The results for this particular case can be generalized to any other case of externalization of constraints. First, because to be *external* implies to be physically materialized. Second, because the constraints of the physical material limit cognitive behavior, and not the other way around. Therefore, to say that a representation is external in respect to some constraints already implies that these constraints are S-dependent, and that we are dealing with iconic semiosis.

To identify the role of abduction in this process, we stress the inferential activity involved in making each move in the game. To solve the game, the player must arrive at some conclusion as how to arrive at a goal state departing from an initial state. To do that, he/she passes through intermediary problem states. The player is making inferences whenever he makes decision as how to pass from one problem state to another. To go from one problem state to another, the player needs to move according to the rules. The rules give the player a certain number of possibilities that he can choose between. This inference is abductive because it is fallible (i.e., it doesn't necessarily conclude the best solution to play) and takes the form of the formation and selection of possible hypothesis of play by departing from a set of constraints.

Figure 5 shows three diagrams depicting constraints in the game. Each node of the diagrams is a problem state, i.e., a particular arrangement of pieces in their places. Each line of the diagrams is a possibility to move from one problem state to another, i.e., to move a piece in the game, according to the rules. One of the nodes is the initial state. Another node is the goal state. To play the game is to go from the initial state node to the goal state node through the possibilities offered by the lines. In the first diagram we have the possible moves as constrained only by the rule 1. In the second diagram we have the same, but now for rules 1, 2 and 3. Let's imagine that these diagrams corresponded to externalized isomorphic representations of the TOH. The first diagram would be a representation in which only rule 1

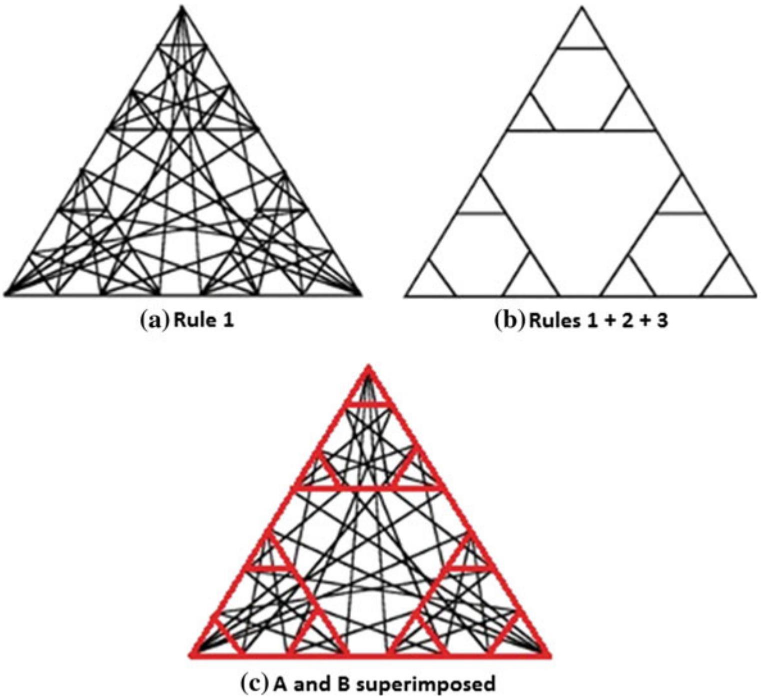


Fig. 5 Constraints of the game for Rule 1 (**a**) and Rules 1 + 2 + 3 (**b**). **c** A superimposition of **b** upon **a**. Adapted from Zhang and Norman [6]

is externalized. The second diagram would be a representation in which all the three rules are externalized. In the game, to perform a move that is out of the rules is considered an error. Therefore, the second diagram, which includes the constraints of all the rules, represents an error-proof scenario (regarding errors that are caused because of moves that are out of the rules). The third diagram shows a comparison between the two isomorphs. In black, is all that was wrong and have been ruled out by the second isomorph in relation to the first. In this sense, we can see the material as a selector of possibilities of play.

A more externally constrained representation is also one where there are fewer possibilities to move the pieces. This doesn't mean that no inferences are present. There is an inferential and perceptual process in the act itself of dealing with the external constraints. For example, when a player chooses to move a cup of coffee to a certain place instead of another because in this better place the coffee will not spill. This inference is supported by external constraints that, as we have seen, are icons of the formal structure of the game. Externalization of constraints (and therefore iconicity) acts as a way to build *better* materials of play. *Better*, here, refers to an economy of possibilities, to the supporting of abductions by the materials of play. In this sense, we have an example of abductive process which is distributed in iconic-embedded features of an externalized semiosis.

7 Situated Semiotic Theory of Mind: Some Implications of Abduction and Iconicity

We have presented an externalist semiotic perspective of cognition, where mind is the result of manipulation of signs and (i) manipulation is described by irreducible forms of inferences; (ii) signs are classified by different morphologies. Abduction and iconicity correspond respectively to the categories of inference and sign processes in which the situated aspect of Peirce's conception of mind is especially conceptualized. Abduction is a weak form of inference (see [23]) related to perceptual features, while the icon is the S-dependent semiotic process. This treatment suggests that a reconsideration of the embodied-situated paradigm's own philosophical foundations can behave in semiotic terms. Peirce's semiotic theory of mind neither restricts representations to symbolic semiosis and inferential processes to deduction and induction as in ortodox representationalism, nor rejects representations and inferences as in anti-representationalism (see Table 4).

This position was exemplified in the case of externalization of constraints in the Tower of Hanoi puzzle. In the example, the task of deciding how to move the pieces of the puzzle was crucially dependent on the materiality of the play, so that isomorphic representations that varied their representational features had great influence on the cognitive behavior of the players (Representational Effect). The game play was facilitated when constraints (the set of rules) were externalized. Externalization of constraints in this context corresponds to the embedment, in an external sign, of better chances to reach an adequate conclusion. We have argued that this process is abductive: it limits the universe of possible moves to a few optimal ones, performing a selection of hypotheses; it provides, through perception, an optimal hypothesis for further consideration; it gives the first step for the solution of the problem.

Table 4 Comparison between orthodox representationalism, anti-representationalism and the Semiotic Theory of Mind

	Representationalism	Anti-representationalism	Semiotic theory of mind
Signs	Symbolic	No	Not only symbolic but indexical and iconic
Inferences	Deductive, inductive	No	Deductive and inductive and abductive
Locus	Internal	External	Inference relies on internal and external resources

8 Conclusion

Recently, the distributed cognition and extended mind approach (see [24, 25]) have questioned the legitimacy of skin and skull to serve as criteria for the demarcation of the boundaries between mind and the outside world. The acceptance of external representation as parts of human cognition leads to different conceptions on the relation between cognition and environment. As we adapt the environment to facilitate our purposes, deploying our mind in external representations, we participate in the construction of cognitive niches, which fundamentally alter our cognitive capabilities (see [26]).

According to Peirce's semiotic theory of mind, thinking *is* semiosis, the process of sign action. While "representationalist", the semiotic theory of mind expands the understanding of signs and inferences beyond orthodox representationalist notions, making it possible to combine representations with an externalist view of the mind. Against any form of internalism, Peirce can be considered a precursor of situated mind and distributed cognition thesis. In the example treated, some of the best solutions, or "ideas" about how to win the game, were embedded in the outside world. Inferences were drawn based on perceptual qualities of material objects rather than an abstract understanding or the 'mind's-eye'. Peirce's broad ideas concerning signs and inferences are an important tool for advancing in the development of an externalist theory of mind.

References

1. Dreyfus, H.L.: Intelligence without representation: Merleau-Ponty's critique of mental representation. Phenomenol. Cogn. Sci. **1**, 367–383 (2002)
2. Wheeler, M.: Reconstructing the Cognitive World—The Next Step. The MIT Press, Cambridge (2005)
3. Chemero, A.: Radical Embodied Cognitive Science. The MIT Press, Cambridge (2009)
4. Skagestad, P.: Peirce's semeiotic model of the mind. In: Misak, C. (ed.) The Cambridge Companion to Peirce, pp. 241–256. Cambridge University Press, Cambridge (2004)
5. Gallagher, S.: Philosophical antecedents of situated cognition. In: Robins, P., Aydele, M. (eds.) The Cambridge Handbook of Situated Cognition, pp. 35–54. Cambridge University Press, Cambridge (2009)
6. Zhang, J., Norman, D.A.: Representations in distributed cognitive tasks. Cogn. Sci. **18**, 1–34 (1994)
7. Ransdell, J.: Some leading ideas of Peirce's semiotic. Semiotica **19**, 157–178 (1977)
8. Hausman, C.R., Charles, S.: Peirce's Evolutionary Philosophy. Cambridge University Press, Cambridge (1993)
9. De Tienne, A.: Learning qua semiosis. Semiot. Evol. Energ. Dev. **3**, 37–53 (2003)
10. Bergman, M.: Reflections on the role of the communicative sign in semeiotic. Trans. Charles S. Peirce Soc. **36**, 225–254 (2000)
11. Queiroz, J., El-Hani, C.: Semiosis as an emergent process. Trans. C.S. Peirce Soc. **42**(1), 78–116 (2006)
12. Savan, D.: An Introduction to C. S. Peirce's Full System of Semiotic. Monograph Series of the Toronto Semiotic Circle, vol. 1, Victoria College, Toronto (1987)

13. Queiroz, J.: Complexification. In: Favareau, D., Cobley, P., Kull, K. (orgs.) A More Developed Sign—Interpreting the Work of Jesper Hoffmeyer, pp. 67–70. Tartu University Press, Tartu (2012a)
14. Queiroz, J.: Dicent symbols in non-human semiotic processes. Biosemiotics **5**, 1–11 (2012b)
15. Stjernfelt, F.: Diagrammatology—An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics. Springer, Dordrecht (2007)
16. Hookway, C.: Truth, Rationality, and Pragmatism: Themes from Peirce. Oxford University Press, Oxford (2002)
17. Hoffmann, M.: Problems with Peirce's concept of abduction. Found. Sci. **4**, 271–305 (2000)
18. Paavola, S.: Peircean abduction: instinct or inference? Semiotica **153**(1/4), 131–154 (2005)
19. Magnani, L.: An abductive theory of scientific reasoning. Semiotica **153**(1/4), 261–286 (2005)
20. Paavola, S.: Diagrams, iconicity, and abductive discovery. Semiotica **186**(1/4), 297–314 (2011)
21. Zhang, J.: The nature of external representations in problem solving. Cogn. Sci. **21**, 179–217 (1997)
22. Chuah, J., Zhang, J., Johnson, T.: The representational effect in complex systems: a distributed cognition approach. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society. Erlbaum, pp. 633–638 (2000)
23. Paavola, S.: Abduction through grammar, critic, and methodeutic. Trans. Charles S. Peirce Soc. **40**, 245–270 (2004)
24. Clark, A., Chalmers, D.: The extended mind. Analysis **58**, 7–19 (1998)
25. Clark, A.: Being There: Putting Brain, Body, and World Together Again. A Bradford Book, Cambridge (1998)
26. Clark, A.: Memento's revenge: the extended mind, extended. In: Menary, R. (ed.) Objections and Replies to the Extended Mind, pp. 1–43. Ashgate, Oxford (2006)

A Logical Model of Peirce's Abduction as Suggested by Various Theories Concerning Unknown Entities

Antonino Drago

Abstract In this paper I will interpret two ways in which Pierce tried to define abduction. The first is Pierce's attempt to find a form of syllogism that would represent abduction. In his last attempts to suggest a syllogism, I show that its conclusion is equivalent to a doubly negated predicate. Since Pierce then considers its corresponding positive predicate, this kind of change of the predicate is called the Pierce principle, which is similar, but weaker than Markov's. The second way is to provide definitions of abduction as a process of construction of an entire theory rather than as the construction of a single law. These definitions are compared with the model of organization of a scientific theory that my previous works recognized through a comparative analysis of many important scientific theories whose aim is to discover a solution of a fundamental problem, how investigate an unknown or ill-defined entity. Each of these theories proceeds by means of reasoning whose conclusion is a doubly negated predicate, which is then changed by the author into the corresponding affirmative predicate, which in turn takes on the role of a hypothesis from which all possible consequences are deduced and then verified experimentally. I will show that this change in the predicate is regulated by the principle of sufficient reason, which is stronger than Peirce's principle. In conclusion, Peirce's attempts to define abduction proved to be intuitive, but imprecise, illustrations of the process of construction of a theory belonging to a particular class of scientific theories making use of non-classical logic.

"It must be confessed that we students of the science of the modern methods are as yet but a voice crying in the wilderness, and saying prepare ye the way for this lord of the sciences which is to come." (7.63) But "I am as confident as I am of death that Logic will thereafter be infinitely superior to what it is as I leave it; but my labours will have done good work towards its improvement" (2.198).

A. Drago (✉)
University of Pisa, Pisa, Italy
e-mail: drago@unina.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_18, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction to Peirce's Notion of Abduction

Since Reichenbach's suggestion [1], most philosophers of science distinguish the context of justification (usually illustrated in deductive way) from the context of discovery (including all elements exploited by a scientist to obtain a new result). But only a small minority of scholars attribute the same importance to the two scientific contexts, and very few scholars consider the latter one as a proper subject of the logical and philosophical investigation of science, because in the majority's view the study of the discovery of hypotheses belongs to psychology rather than to the scientific method. Moreover, the widespread and rooted "dogma" that all valid scientific theories are deductively organised from principles-axioms leads almost all scholars to assume that there is no place for non-deductive processes in a complete scientific theory. As a consequence, the teaching of scientific theories is at present confined to transmitting deductive thinking only, without any reference to non-deductive reasoning.

On the other hand, Peirce maintained that logicians are too concerned with "[the] security (approach to certainty) of each kind of reasoning" (8.384), to devote due attention to non-deductive reasoning".[1] McMullin remarked that, against the dominant positivistic attitude of his times, Peirce

> had no hesitation in inferring [non-deductively] to unobservable entities…. "Attempts to explain phenomenally given elements as products of deep-lying entities" (using molecules to explain heat is his example) are [according to Peirce] entirely legitimate; in fact, this phrase may be said to describe "as well as loose language can, [nothing less than] the general character of [the search for] scientific hypothesis [i.e. abduction]. (8.60)" [3, p. 86].

Hence, Peirce explored an unusual theme of philosophy of science; he called himself "an explorer of untrodden ground" (2.102). In fact, he devoted much time to classifying arguments used by scientific works (2.461). The result was, instead of the common view of two processes of induction and deduction, three more or less independent inference processes:

> If we are to give the names of Deduction, Induction and Abduction to the three grand classes of inference, then Deduction must cover every attempt at mathematical demonstration, whether it is to relate to single concurrences or to "probabilities", that is to statistical ratios; Induction must mean the operation that induces an assent, with or without quantitative modifications, to a proposition already put forward, this assent or modified assent being regarded as the provisional result of a method that must ultimately bring the truth to light; while abduction must cover all the operations by which theories and conceptions are engendered (5.590).[2]

This quotation makes apparent that Peirce, while he attributed the usual meaning to deduction, attributed special meanings to the other two notions.

[1] For an introduction to Peirce's thinking on inference processes, see (Fann). Two numbers in round brackets refer to [2], the first number means the volume and the second number the section.

[2] In the following I disregard the statistical case because, in my opinion, it is less interesting.

Indeed, according to Max Black, there exists no satisfactory theory concerning the last two processes [4, p. 169].[3] They

> cover all cases of non-demonstrative argument in which the truth of the premises, while not entailing the truth of the conclusions, purports to be a good reason for belief in it. Such arguments may be also called "ampliative", as C.S. Peirce called them, because the conclusion may presuppose the existence of individuals whose existence is not presupposed by the premises [4, p. 169].

On his side, Peirce conceived induction in a novel way:

> Induction is the experimental testing of a theory… It sets out with a theory and it measures the degree of concordance of that theory with facts (5.145).
>
> Induction consists in [the final process of] starting from a theory [already constructed], deducing from it predictions of phenomena, and observing those phenomena in order to see *how nearly* they agree with the theory" (5.170).

Even more novel is Peirce's notion of abduction.

> In very many questions, the situation before us is this: We shall do better to abandon the whole attempt to learn the truth, however urgent may be our need of ascertain it, unless we can trust to the human mind's having such a power of guessing right that before very many hypotheses shall have been tried, intelligent guessing may be expected to lead us to the one which will be supported by all tests, leaving the vast majority of possible hypotheses unexamined (6.530).

> Abduction is the process of forming an explanatory hypothesis (5.171).
>
> [Abduction] is the provisional adoption of a [testable] hypothesis (1.68).
>
> I reckon it as a form of inference, however problematic the [obtained] hypothesis maybe held (7.202–7.207).
>
> [It] only infers a *may-be* (5.189).
>
> Suppose something of a different kind from what we have directly observed, and frequently something which it would be impossible for us to observe directly (2.640).
>
> [Abduction brings] a wholly new element into the conclusion (5.192).
>
> By hypothesis [=abduction] we conclude the existence of a fact quite different from anything observed, from which, according to known laws, something observed would necessarily result (2.636).
>
> As a general rule, hypothesis is a weak kind of argument. It often inclines our judgement so slightly toward its conclusion that we cannot say we believe the latter to be true; we only surmise it may be so (2.625).
>
> The conclusion of an abduction is problematic or conjectural, but is not necessarily at the weakest grade of surmise, and what we call assertoric judgements are, accurately, problematic judgements of a high degree of hopefulness (5.192).
>
> The conclusion of Hypothetic Inferences cannot be arrived at inductively, because their truth is not susceptible of direct observation in single case. Nor can be conclusions of Inductions, on account of their generality, be reached by hypothetic inference (2.714).
>
> [Abduction] cannot be admitted as a hypothesis, unless it be supposed that it would account for the facts or some of them (5.189).

[3] Feferman [5, pp. 77–93] summarises Lakatos's attempt to introduce a new way of considering non-deductive reasoning in science. His appraisal of Lakatos is that real novelty is yet to come.

2 Representing Abduction by Means of Syllogisms

However, Peirce staunchly emphasised that abduction is reasoning and also formal reasoning. According to Peirce, "The successful theories are not pure guesses, but are guided by reasons" (2.638). Indeed, he stressed that an abduction, which generates a theory, is not a merely intellectual jump or guess, but it is "… *reasoning* and though its *security* is low, its *uberty* [=fecundity] is high" (8.388).

> Although it is very little hampered by logical rules, nevertheless is logical inference, asserting its conclusion only problematically or conjecturally, it is true, but nevertheless having a perfect definite logical form (5.188).

Peirce tried for a long time to translate his ideas into syllogisms of a similar kind of Aristotle's.

The first instance of Peirce's translation of abduction in a kind of syllogism was the following one: "All balls in this urn are red; all balls in this particular random sample are red; therefore, all balls in this particular random sample are taken from this urn… It should be clear that abduction is never necessary inference."[4] But unfortunately it was contested.[5]

A further instance was the following one:

> The surprising fact C is observed,
> But if A is true, C would be a matter of course;
> Hence there is a reason to suspect that A is true (5.189).

Surely, the common view of classical logic does not support the above statements. Hintikka states that abduction is a "puzzle" in the sense "… it still is extremely difficult to see why abduction whatever it is or may be, can be not only a rational operation but even a logical inference, in any sense of logical inference" [9, p. 504].

In literature, as Hoffmann remarks, there exist two strategies for interpreting Peirce's "logicality" of abduction, i.e. either a "heuristic strategy" or to claim that it is a matter of "practical character" alone [10, p. 277].

[4] Let us recall that already Aristotle (Aristotle *Topics* b, p. 29ff.) considered "a syllogistic inference from a major premiss, which is certain, and a merely probable minor premiss, to a merely probable conclusion." A clever re-construction of the various kinds of syllogism Peirce suggested is in [6, pp. 436–439]. In his opinion the instance in the following section represents the more general schema.

[5] See, beyond [7] the new definition that Cellucci [8, p. 235] suggested: "Given a set of sentences Γ and a sentence C not derivable from Γ, to find a sentence A such that C is derivable from Γ + A. The set Γ consists of the already available hypotheses, or background hypotheses, C is the problem to be solved, A is the new hypothesis sought for."

But let us recall that some scholars introduced into the history of logic a distinction between two traditions, i.e. "Logic as Calculus and Logic as Language",[6] whose origin can be traced back to Aristotle and beyond.

> With Frege logic became a universal language, while Boole's "universal classes" and de Morgan's "universe" of discourse are contexts that can be changed at will… For Frege it cannot be a question of changing universes…. In a similar way Hintikka distinguishes "*language as the universal medium* (or the *universality of language*) and *language as calculus* (or *the model-theoretical view of language*), where language is like a calculus in that it can be freely assumed a new interpretation [10].

According to Hintikka [13] Peirce belongs to the latter tradition owing to his

> … greater emphasis on "operation of inference" than on "a universal language of mathematicians". For Peirce choosing a "universe of discourse" is the first step in representing logical relations. [10, p. 276].[7]

3 A Peirce Formal Principle Derived from a Peirce Syllogistic Instance of an Abduction

Now let us consider the latter syllogism which Peirce suggested as instantiating an abduction. Psillos illustrates it:

> It involves generation of some hypothesis A with excess content in virtue of which the explanandum C is accounted for, where the explanatory connection between A and C is captured by the counterfactual conditional "If A is true, C would be matter of course". But is also allows the detachment of the antecedent A from the conditional and hence its acceptance in its own right. The detachment of the antecedent A requires reasons and these are offered by the explanatory connection there is between the antecedent and the consequent [16, p. 133].

But Psillos dismisses Peirce's word in the conclusion, "suspect" (7.220); it is not a certainty from which one is allowed to detach the antecedent—an operation which is legitimate in classical logic from a certain result only.

I suggest that the word "suspect" substitues a doubly negated statement: "It is not the case that it is not…",[8] or "We cannot exclude that…". These doubly negated statements are not equivalent to the corresponding affirmative statements, since the latter are not supported by scientific evidence (=DNSs).

[6] Heijenoort [11]. See also [12, fn. 18].

[7] Of course, also every supporter of an alternative kind of logic stressed that his logic does not belong to the dominant tradition. For instance, see [14] for intuitionist logic and [15] for modal logic.

[8] Here and in the following, emphasis is added to make clear to the reader the two negations included in a DNS; modal words are underlined too, being equivalent to DNSs; e.g. possible = it is not the case that is not. More in general, it is well-known that the modal logic is translated into intuitionist logic by means of S4 model.

This point has to be clarified, because the current usage of the English language exorcises DNSs as representing a characteristic feature of a primitive language,[9] while the following three well-known DNSs belonging to mathematics, physics and classical chemistry show that this logical feature pertains to scientific research since its origin.

In Mathematics it is usual to state that a theory is "without contradictions"; to state the corresponding affirmative statement, i.e. the consistency of the theory, is impossible, owing to Goedel's theorems. In theoretical physics it is usual to study the in-variants of a theory, which does not mean that the in-variant magnitude is remained fixed. In order to solve the problem of what the elements of matter are, Lavoisier suggested defining these unknown entities by means of a DNS: "If… we link to the name of elements or principles of corps the idea of last term to which [through chemical reactions] arrive at the analysis, all the substances which we were not capable to decompose through any tool are for us, elements",[10] where the word 'decomposable' naturally carries a negative meaning and stands for 'non-ultimate' or 'non-simple', which does not mean "simple" [22, p. 7].

According to mathematical logic, the failure of the double negation law in a sentence qualifies this sentence as belonging to non-classical logic—in particular, to intuitionist logic [23, 24]. As a matter of fact, Grzegorczyk proved independently that scientific research may be formalized through statements belonging to intuitionist logic [25].

I conclude that also Pearce's syllogism belongs to intuitionist logic.[11] In formal terms the third statement is written as.

$$\exists x \neg\neg f(x)$$

Hence, Peirce conceived the formal representation of abduction according to an intuitionist predicate, which is not equivalent to $\exists x f(x)$.

The conclusion drawn by Psillos is in fact what Peirce himself wanted to obtain in order to test an affirmative statement; indeed, it could not be otherwise, since is not possible to test a DNS. Hence, Peirce implicitly changed the above predicate into the corresponding affirmative predicate:

[9] English linguists are dominated by a long tradition which L. Horn called a "dogma" [17, pp. 79ff.; 18]. This linguistic dogma asserts the absolute validity of the double negation law: whenever a DNS is found in a text, it has to be changed into the corresponding affirmative statement, because those who speak by means of DNSs want to be, for instance, unclear. Evidence for this "dogma" is the very small number of studies on double negations in comparison with the innumerable studies on single negation. The historians of science ignored DNSs; Klein [19] suspected the relevance of the linguistic expressions of the scientists. No wonder Peirce did not notice this logical aspect, which moreover pertains—as we will see—to particular scientific theories only.

[10] The numerous DNSs of his texts have been listed by my papers [20, 21].

[11] Also the conclusions of the forms 2.702 and 2.706 are non-affirmative statements: "… probably and approximately….".

$$\exists x \neg\neg f(x) = \; > \exists x\, f(x)$$

By examination of Dummett's table of the relations of implication between any two predicates of intuitionist calculus [23, p. 29], we see that the above doubly negated predicate is not equivalent to the antecedent of the well-known Markov's principle [23, p. 21; 26, pp. 27 and 129].

We can call this change the *Peirce principle*, since it is not a logical deduction, but a change of an intuitionist predicate into the corresponding classical predicate.

Result no. 1: *Last Peirce's instance of a syllogism illustrating an abduction suggests a logical change—here called Peirce's principle—from an intuitionist to its corresponding a classical predicate.*

4 Abduction as Theory Generation

Sections 1 and 2 illustrated the best hints offered by Peirce about his idea of abduction. Unfortunately, about the distinction between abduction and induction, Peirce wrote in 1910: "… in almost everything I printed before the beginning of this century I more or less mixed up hypotheses [abduction] and induction" (8.227). Moreover, when suggesting the main features of abduction, Peirce implicitly referred to his wide knowledge of scientific theories; but he did not offer certain instances of scientific theories to which his notion refers[12]; they would have severely limited the great variety of later interpretations of his notion.

One scholar wrote: "Over many years Peirce modified his views on the three types of argument, sometimes changing his views but mostly extending them by expanding his commentary upon the original trichotomy" [29, sect. "Induction, Abduction, Deduction"]. Hence, at least two periods have to be distinguished in his reflection on abduction, which spanned more than fifty years [30, pp. 648ff] McMullin adds: "It is not easy to disentangle the theme abduction/retroduction from the enormously complex and sometimes idiosyncratic metaphysical and psychological system Peirce labored to build and rebuild." [3].

No surprise if the abduction process received a variety of interpretations, as Douven states:

> Most philosophers agree that this type of inference is frequently employed, in some form or other, both in everyday and in scientific reasoning. However, the exact form as well as the normative status of abduction are still matters of controversy.… Precise statements of what abduction amounts to are rare in the literature on abduction. (Peirce did propose an at least fairly precise statement but… it does not capture what most nowadays understand by

[12] Apart two instances: (1) Kepler's theory of Mars' orbit (1.73); which was further illustrated by (Hanson, ch. IV). But see the valid criticisms in [27]. In any case, Kepler's theory belongs to a too informal context for suggesting certain conclusions on his underlying logic. (2) The kinetic theory of gas (2.639) which however, is too loosely treated by Pearce; moreover, also this instance was contested [28].

> abduction). Its core idea is often said to be that explanatory considerations have confirmation-theoretic import, or that explanatory success is a (not necessarily unfailing) mark of truth [31; Introduction and beginning of Sect. 2].[13]

Thus, when reading Peirce's writings one is forced to interpret them; my quotations represent no more than a particular selection of his words, whose merit is to offer a more precise definition of abduction.

In general, Peirce's writings leave room for two interpretations, i.e. abduction as the selection of a hypothesis among several others on the basis of experimental data, or as an argument about whether a hypothesis is plausible for further theoretical and experimental developments.[14] The former case was intensively analysed, e.g. by [38], but was contested by Hintikka [9, pp. 506–511].

Let us consider the latter case—the "making a hypothesis" (2.623). McMullin suggested a distinction which in my opinion is very important. He contrasted the instrumentalist account of explanation, which reduces an entire theory searching for new hypotheses to a complicated form of induction, i.e. a theory to a possible law [3, p. 95]. At present the great variety of scientific theories shows that one has to distinguish two levels of a theory, i.e. the single law—which is what interests the instrumentalists—and the totality of a theory composed of several laws, theoretical terms, principles and mathematical techniques [3, p. 90]:

> Induction is *strictly* limited to the observable domain. And it is only in a very weak sense explanatory. Laws may explain singular occurrences, of the sort that the D-N model was devised to handle. But these are the material of history and engineering, not of natural sciences as chemistry of physics. Laws are the *explananda*; they are the questions, not the answer.
>
> To explain a law, one does not simply have recourse to a higher law from which the original law can be deduced. One calls instead upon a *theory*, using this term in a specific and restricted sense. Taking the observed regularity as effect, one seeks by abduction a causal hypothesis which will explain the regularity. To explain why a particular sort of thing acts in a particular way, one postulates an underlying structure of entities, processes, relationships, which would account for such a regularity. What is ampliative about this, what enables one to speak of this as a strong form of understanding, is that if successful, it opens up a domain that was previously unknown, or less known.
>
> The causal inference here is therefore not the abduction alone, which is still a conjecture, even if a plausible conjecture. It is the entire process of abduction, deduction, observational testing, and whatever else goes into the complex procedure of theory appraisal [3, pp. 90–91].[15]

[13] See also [32, 31, "1.2 The ubiquity of abduction"; 33, pp. 448–449; 34]. According to [35], the two kinds of non-deductive reasoning receive, at the present time, two main interpretations, one mainly in Philosophy, the other mainly in Artificial Intelligence.

[14] This distinction was introduced by [36, pp. 85–92]. See also [37, pp. 41–44].

[15] See also [39]. McMullin calls "retroduction" the entire process including the three stages of abduction, deduction and induction. I call it the "development of a theory looking for knowledge of an unknown entity", or even the "organization" of such a theory. Whereas McMullin does not instantiate the theories in which retroduction works, abduction in the sense I mean concerns only one particular kind of scientific theory.

Since the development of a scientific theory links together several kinds of statements in a systematic context which has to be considered also as a whole, the subject of abduction requires more than simply reflecting on single statements—either a law or a mere hypotheses of law -, as scholars studying inference processes usually do. In sum, while Peirce's induction refers to an experimental law, Peirce's abduction refers to *an entire theory*. In McMullin's words: "The product of retroduction [=abduction] is *theory* or causal explanation. It is distinct from empirical *law*, the product of the simpler procedure of induction… It is a question of the amplitude of our world." [3, pp. 93, 95; emphasis in the text].

Unfortunately, Pierce sometimes seems to think of a theory as represented by a single formula; several of Peirce's statements support the above idea. E.g.:

> The purpose of a theory may be said to be to embrace the manifold of observed facts in one statement, and other things being equal, the theory that best fulfils its function is the one that brings the most facts under a single formula" (7.410).
>
> [In other words, a theory has] to substitute a great series of predicates forming no unity in themselves for a single one which involves them all (5.276).
>
> Abduction seeks a *theory*. Induction seeks for facts. In abduction the consideration of facts leads to a hypothesis. In the induction the study of the hypothesis suggests the experiments that bring to light the very facts to which the hypothesis had pointed" (7.218).

The last quotation clearly shows that Peirce shifts between a "theory" and "hypothesis" of a single law[16];

However, he often links the notion of abduction precisely to a theory (in the following I always add emphasis to the word "theory"):

> Abduction in the sense I give the word, is any reasoning of a large class of which the provisional adoption of [no more than] an explanatory hypothesis is the type. But it includes processes of thought which lead only to the suggestion of questions to be considered, and includes *much besides* (2.544 note; emphasis added).
>
> [It] "covers all the operations by which *theories* and conceptions are engendered" (5.590).
>
> [Abduction] consists in examining a mass of facts and in allowing these facts to suggest a *theory*" (8.209).
>
> Abduction consists in studying facts and devising a *theory* to explain them. Its only justification is that if we are ever to understand things at all, it must be in that way (5.145).
>
> Broadly speaking, abduction covers all the operations by which *theories* and conceptions are engendered (5.590).

His words provide even more support for this interpretation of abduction as theory generation, when he compares the three inferential processes:

[16] Fann remarks that in a first period of time "Peirce conceived both induction and hypothesis as species of "reduction of a manifold to unity" (5.276). … The function of hypothesis is "to substitute for a great series of predicates forming no unity in themselves, a single one which involves them all." It is therefore, like induction, a reduction of a manifold to unity." [37]. In this description one can see the achievement of both a law and a theory; however, it is the latter that mainly represents the unity of a field of phenomena.

> We naturally conceive of science as having three tasks—(1) the discovery of Laws, which is accomplished by induction; (2) the discovery of Causes, which is accomplished by hypothetic inference [=abduction]; and (3) the production of Effects, which is accomplished by deduction (2.713–2.714).
>
> [Moreover, abduction] furnishes the reasoner with the problematic *theory* which [then] induction verifies" (2.624).
>
> [Induction] sets out with a theory and it measures the degree of concordance of that theory with fact. It never can originate any idea whatever. No more than deduction. All the ideas of science come to it by the way of Abduction. Abduction consists in studying facts and devising a *theory* to explain them. Its only justification is that if we are ever to understand things at all, it must be in that way. Abductive and inductive reasoning are utterly irreducible, either to the other or to Deduction, or Deduction to either of them… (5.146).
>
> Induction shows that something actually is operative, Abduction merely suggests that something *may be* (5.171).

As McMullin puts it:

> Abduction begins from the facts without having any particular theory in view, motivated only "by the feeling that a theory is needed to explain the surprising facts". Induction, on the other hand, begins from an hypothesis already arrived and seeks for facts to support that hypothesis [3, p. 88].[17]

In other words, abduction is the preparatory step of a theory, deduction the intermediate step for obtaining theoretical predictions and induction is the final phase of the inquiry into the validity of the latter ones.

In the following I will study the logic of abduction as the creative generation of a new theory.

5 The Kind of Abduction Presented by the Original Texts of some Scientific Theories

In order to throw light on the above problems, I will exploit the wide variety of theories in the history of hard science.

By following Peirce's suggestion for investigating theories concerning "unrecognised objects", or "surprising facts",[18] I select among past scientific theories those theories that inquired into entities which at their respective times were either unknown or ill-defined. I will disregard their presentations in current textbooks, where most of them conform to a deductive model of organisation, i.e. to a context of justification; rather I will refer to the original texts in which they have been presented.

[17] See also [29].

[18] According to Peirce, "The [new] hypothesis must explain the surprising fact" (5.189); in my opinion, the "surprising fact" has to be seen as a non-paradigmatic fact, i.e., in Kuhn's terms, an anomaly; whose solution generates a new theory which eventually obliges scientists to revise the basic tenets of past paradigm.

I list the 10 more relevant theories of this kind: Lavoisier's theory of the elements of matter (1789), Avogadro's theory of atoms (1811), S. Carnot's theory of the best machine for converting heat into work (1824), Lobachevsky's theory regarding non-Euclidean parallel lines (1826), Einstein's theory of the invariant mechanics under the Lorentz group (1905), Einstein's theory of quanta (1905), Planck's theory of quanta (1913), Kolmogorov's theory of intuitionist formal logic (1925), Church's theory of computable functions (1936), Markov's theory of constructive real numbers (1964). Almost all them are currently recognised as "revolutionary" theories, whose importance in the history of science cannot be underestimated.

Each of these theories starts from the problem of how to arrive at the knowledge of an unknown entity which was "surprising" since at that time it was intractable by means of the current scientific theories and techniques; each founder of one of the above theories wants to discover a hypothesis regarding the unknown entity; and then he reasons about it in order to make it plausible enough for deductively obtaining consequences to be tested; whose positive answers eventually solve the given problem. Hence, each one, taken as a whole, constituted a non-deductive process of inference since it moved from disparate data to a plausible hypothesis of a universal nature. From this investigation we conclude that there exists a variety of important scientific theories which are candidates for representing abductive processes of inference. I call the organisation of these scientific theories a problem-based one (PO).

An inspection of their original texts shows that they make use of DNSs; in few pages of each text more than 50 DNSs occur; moreover, these DNSs play crucial roles in the presentations of scientific theories. Usually, their mere sequence is enough to provide the logical thread of the entire first part of the theory; in other words, the sequence of DNSs represents the core of the author's reasoning, precisely as Grzegorczyk's interpretation of scientific research suggests. In other words, the development of each of the above-mentioned theories proceeds through a meaningful sequence of DNSs. Hence, the authors of the above-mentioned theories using DNSs—surely, by mere ingeniousness—presented their theories in non-classical logic, above all intuitionist logic.[19]

Let us recall that Peirce's hypothesis cannot be excluded a priori, or in other words is merely likely (notice the emphases and their above illustrated meanings in footnote 8), In fact, due to the lack of certainty, its content is appropriately expressed precisely by means of a DNS. Hence also in a scientific theory, Peirce's abductive process and in particular the resulting hypothesis have to be expressed by means of DNSs.

[19] Actually, that holds true in the first parts of each text only. What is the first part will be defined in the next section. In some texts DNSs occur also in the latter part too, but they are few in number and do not play an essential role in the development of the theory. The above mentioned theories (except for Church's theory) have been analysed in their FDNs by the following papers [20, 21, 41–47]. Being impossible to present all these DNSs, in the next section the list of only the most relevant DNSs of the above-mentioned theories will be presented.

Result no. 2: Peirce's abduction, which aims at invent a hypothesis that engenders a new theory, has to be conceived by means of FDNs and more in general in the framework of non-classical logic.

Furthermore, the texts of the above theories provide evidence for a well-defined common way of reasoning, because a chain of *ad absurdum* arguments characterises each of them—apart the writings by Lavoisier [22], Einstein's special relativity [48] and Markov's theory of constructive numbers (Markov), which lack of this kind of argument.[20] The DNSs produce informative logical arguments by composing *ad absurdum* arguments, the conclusion of each of them is a DNS, which then plays the role of a premise for the next argument. The most well-known instance of this kind of argumentation is presented by Sadi Carnot's theory of thermodynamics; in the original text his main argument, i.e. the *ad absurdum* theorem about the efficiency of heat engines, repeated in most current textbooks of thermodynamics is the last one of a chain of other *ad absurdum* arguments.

All that indicates that the above-mentioned authors reasoned in non-classical logic. It is not possible to underemphasise the fact that these authors followed this formal way of arguing by mere ingenuousness.

Due to trivial historical reasons Peirce did not have the concrete possibility to refer to both DNSs and non-classical logic as presented by the above theories.[21]

6 The Common Logical Predicate Concluding the above Theories and its Translation into a New Hypothesis

In a previous paper [44]. I listed the conclusions of all these logical developments. For instance:

——S. Carnot: "… if it were possible in some way to make caloric yield a greater amount of motive power than was produced in our first sequence of operations… it is inadmissible". ([50, pp. 20–21].[22]

[20] Einstein's paper on special relativity presents an universal idea: "We cannot attribute to the notion of simultaneity an absolute [=non relative, all results of physical measurements being relative] meaning" [48, p. 891], but its elaboration till up to a plausible hypothesis is conducted by operative means (rods and clocks, plus mathematics), not through *ad absurdum* arguments. The aim of Markov's paper [49] was to discover a general algorithmic way for constructing all constructive real numbers; to this aim his theory refers to a general theory of algorithms where he assumed a specific principle which allows him to argue without *ad absurdum* arguments. See the paper [44], whose Appendix lists Markov's DNSs.

[21] However, two facts show a progressive refinement of his logical arguments towards intuitionist logic. In 2.623 he offered instances of abduction whose conclusions are affirmative statements, hence of classical logic, while in 7.220 he concludes by means of the word "suspect". His intermediate instances of "statistical deductions" (2.702 and 2.706) may explain this passage.

[22] Carnot's thermodynamics starts with the problem of knowing the maximum efficiency in heat/work conversions; in order to solve it, he looks for a new method by reasoning through DNSs

Lobachevsky: the new hypothesis of two parallel lines "… can likewise be *admitted* without leading to any contradiction in the results" [51, p. 19].[23]

Kolmogorov: "The use of the principle of excluded middle never leads to a contradiction [52, p. 431].[24]

It is a remarkable fact that the above conclusions are formalised by the same logical predicate:

$$\neg \exists x \neg f(x).$$

Some authors reached somewhat different conclusions.

Avogadro: "[All] The quantitative proportions of substances in compounds do not seem to depend on other than both the relative number of molecules which combine and on the number of composed molecules which result" (53, p. 58),[25] i.e. all the proportions among the quantities in the combinations of the substances depend on other than …,

Kleene's statement on Church's thesis: "Every general recursive function cannot conflict with the intuitive notion which is supposed to complete…" [54, pp. 318–319].[26]

All the above statements are formalised by the same following predicate:

$$\forall x \neg \neg f(x).$$

While in non-classical logic the two quantifiers are not interchangeable in general, it is a remarkable fact that the latter predicate is equivalent to the former predicate [23; table of p. 29]. We call it A^I.

Summarising, the first part of a theory aimed at solving a problem is a goal-oriented logical theory aimed at stating through a final indirect proof a universal predicate of the A^I form.

Notice that each conclusion A^I is a mere hypothesis, without any assured certainty, which is precisely what Peirce stated about the result of an abduction

(Footnote 22 continued)
about his celebrated cycle of four transformations. The sequence of DNSs has as a crucial point his well-known *ad absurdum* proof about the efficiency of reversible engines, which is not less than the efficiency of the irreversible engines. Carnot's book was analysed through its DNSs in the paper [47].

[23] This book was analysed through its DNSs in my paper [40]. Lobachevsky dealt with the problem of how much are the parallel lines. He proved through DNSs six theorems, almost all *ad absurdum*, concluding that the hypothesis of two parallel lines is plausible.

[24] Kolmogorov's foundation of intuitionist logic argues systematically by means of DNSs — called by him "pseudotruths"—and an *ad absurdum* proof [52, p. 431] -, stating the above conclusion. An analysis of the paper through its DNSs is given by [40].

[25] Avogadro wanted to discover the last elements of matter through some of their properties, first of all their number in a given volume.

[26] Scholars argued about the problem of what a computable function is by comparing the intuitive notion with several formal notions; all latter notions being equivalent, they arrived at the above DNS, although its affirmative translation was commonly exploited.

process of inference. Yet, each of such hypotheses constitutes the central statement of an entire scientific theory.

Result no. 3: *Peirce's abduction aimed at inventing a new plausible hypothesis generating a new theory has to be conceived as a logical process leading to the statement of a universal predicate of the form A^I in one of its two versions.*

7 The Change in the Kind of Logic

Let us summarise the previously illustrated development of a PO theory. A scientist constructing a PO theory argues by means of DNSs belonging to non-classical logic, in order to prove by an (at least one) *ad absurdum* proof a universal doubly negated predicate. A comparative analysis of the remaining 7 texts shows that at this point in the development of the theories, some of the above-mentioned authors—Avogadro, Sadi Carnot, Lobachevsky, Einstein—overtly declared a radical change.

Avogadro's statement subsequent to the one quoted above is as follows: "It must then be admitted that very simple relations also exist between the volumes of the gaseous substances and the number of simple or compound molecules which form them". It constitutes Avogadro's celebrated law concerning the molecular constitution of any kind of matter.

Immediately after reiterating the conclusion of p. 20, Carnot restates the statement but in the affirmative version: "The various methods of developing motive power can be arrived at either by using different substances or by using the same substance under differing conditions—for example, a gas at two different densities" [50].

After proving proposition no. 22, Lobachevsky wrote: "[The affirmative hypothesis corresponding to my doubly negated predicate introducing two parallel lines *deductively*] *founds* a new geometry…." [51, p. 19; emphasis added].

In the 1905 paper on quanta, Einstein, after stating the above-mentioned DNS applied the affirmative hypothesis corresponding to it: "According to the just obtained result, let us assume… both the incident and emitted light consist of energy quanta of magnitude *Rβv/N*…" [55, p. 372] in order to derive deductions capable of interpreting three new phenomena in the last three sections of the paper.

When introducing his result Kolmogorov states it through the corresponding affirmative statement: "We shall prove that all the finitary conclusions obtained by a transfinite use of the principle of excluded middle are correct…." [52, p. 416].

One has to suppose that the universal nature of the last DNS suggests to an author of a PO theory that the maximum amount of evidence has been obtained for extending his knowledge in a controlled way; he apparently has completed his reasoning through both the uncertain DNSs and the *ad absurdum* proofs; the result obtained in this former part of the theory now appears to him as a plausible hypothesis; hence, it can in the second part of the theory play the role of a principle axiom; to this end it has to be changed into the corresponding affirmative predicate, from which a deductive development then starts.

Formally, the author of a PO theory changes the doubly negated predicate A^I belonging to non-classical logic into the corresponding classical predicate A, belonging to classical logic; in formal terms:

$$\forall x \neg\neg f(x) = \ > \forall x f(x)$$

Let us remark that all scientists of the quoted PO theories, practiced in predicate logic the method of the *ad absurdum* proof by adding a strong qualification; this proof is universal in nature with respect to all the problems involved in the basic problem. One more qualification is suggested by Markov's principle, i.e. to argue on decidable predicates; this qualification is implicit in the other authors, otherwise no scientific conclusions are possible.

The subsequent part of the theory is then deductively developed according to classical logic; in fact, classical reasoning is preferable since it is easier than non-classical, as is recognised when comparing the before/after of the four translations of propositional intuitionist logic into classical [56, p. 57]: in non-classical logic all the following properties of classical logic are not possible: the law of double negation unites the affirmative and the corresponding doubly negated sentences, the law of the excluded middle makes absolutely sure any logical implication, all indirect proofs are converted into direct proofs and moreover the dilemma (a combination of implicative and alternative propositions) imposes a single consequent.

Hence, in the midst of the global development of the theory the author, in order to recover classical logic, changes the kind of logic through the predicate change A^{I} => A[27]; at the same time the problem-based organisation of the theory changes into a deductive organization.

In Peirce's terms, the first part of the theory constitutes an abduction process, moving from uncertain data to a plausible predicate of a universal nature; the second part of the theory begins the process of deductive development from the predicate which has been changed into a classical predicate; the deductions have then to be tested with experimental data.

It is a historical fact that Lobachevsky tested the deductions drawn from the hypothesis obtained by his reasoning, by comparing them with astronomical observations. Also Einstein's two papers of the year 1905 obtained deductively some theoretical results which have been (successfully) compared with experimental phenomena. In general, the novelties of the second part of a theory aimed at solving a problem are tested for validation by testing them with real facts. This process is precisely what Peirce called induction.

[27] Achinstein (Achinstein 1971, p. 123–124) interprets Peirce's abduction as the achieving of a only *plausibile* hypothesis; which then he formalises in Hanson's way [36] as 'All F's are G's".

Result no. 4: *The global development of a theory aimed at solving a problem is composed precisely of the three logical stages that Peirce suggested; i.e. abduction, deduction and induction have been identified.*

Peirce could obtain these results by an accurate examination of some theories preceding him, above all, Lobachevsly's; in fact, he studied non-Euclidean geometries, but at that time no logical analysis of the original texts had been performed. However, Peirce maintained that "each chief step in science has been a lesson in logic" (5.363); in the history of science this statement can be considered to represent both Lobachevsky's and Einstein's theories. This agrees with what Peirce wrote: "Modern methods have created modern science; and this century… has done more to create new methods than any former equal period" (7.61).

8 Leibniz's Principle of Sufficient Reason

Now, let us recall Leibniz's logico-philosophical principle of sufficient reason:

> Two are the principles of the human mind: the principle of non-contradiction and the principle of the sufficient reason…, [which is the following one:] nothing is without reason, or everything has its reason, although we are not always capable of discovering this reason… (Leibniz).

In the latter principle the first statement, "Nothing is without reason", is a DNS; the same Leibniz then explains why it is so: in order to simplify the reasoning, from the previous statement we want to obtain the affirmative sentence "Everything has a reason"; but Leibniz emphasises that we do not always have sufficient evidence for affirming such a reason with certainty; hence, the first statement, not being equivalent to the next one, is a DNS and thus it alone holds true; the second one is a statement without operative evidence, similar to that of the existence of a single parallel line.

Remarkably, in mathematical logic the first part of Leibniz' principle is formalised by the same formula A^I above, like the previous conclusions of PO theories:

$$\neg\exists x \neg f(x)$$

In philosophical terms, Leibniz' statement appears to be the philosophical principle guiding an author of a PO theory in his non-classical reasoning, aimed at concluding the first part of the theory by achieving a logical predicate A^I, which is formally the same as the first part of this principle.[28]

[28] Compare it with Thomson's words when he discovered the electric charge of cathode rays: "I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter." (Thomson; quoted by [57, p. 17]). But also Markov, when stating his principle, is recognised to argue according to PSR° [23, p. 22].

Moreover, let us remark that the above quotation of Leibniz' principle has two parts; the first part consists of a doubly negated predicate, whereas the second part is the corresponding affirmative sentence "… everything has its reason".[29] We can interpret the two parts as a change of the non-classical predicate to the corresponding classical one. Remarkably, also this change is represented by the same formula A^I -> A formalising the final move of a PO theory. (In the following it will be called PSR°). This move is not a matter of a particular kind of logic—be it classical logic or not -, but is a philosophical "principle" of rationality.

All the above authors were unaware of their own way of reasoning; however, the above comparative analysis of their original works showed that all they tried to adjust their way of reasoning to a common pattern. Moreover, the logical strategy of an author of a PO theory may be considered as inspired—or summarised—by Leibniz' PSR°.

9 Peirce's Intuitive Approximation of this Kind of Rationality

Being confined to classical logic, Peirce considered the second part only of PSR°, i.e. the A predicate: "Everything has a reason"; which alone is of course an idealistic statement. He moreover replaced "reason" with a word which allows a deterministic conception of reality, i.e. "cause"; this move agrees with the deductive organisation of a theory, the alternative organisation governed by the PSR°. Moreover, although he knew Leibniz's thought much better than contemporary philosophers [58], he certainly did not agree with Leibniz' principle.[30]

However, Peirce was not so far from this kind of reasoning. Since his purpose was ultimately to define the logic of inquiry, his definition of logic was broader than the usual one: "the art of devising methods of research,—the method of methods" (7.59).

He conceived the reasoning process in a more general way than the usual one of classical logic. In order to escape from its narrow view of reasoning he describes a reasoning that includes several elements; at first glance they seem to concern the psychology of the reasoner, but are meant to represent intuitive logical processes

[29] Notice that also this change is implicitly justified by the following *ad absurdum* argument: "It is absurd to reject the consequent ("Everything has a reason"), otherwise, the reality is not rational".

[30] I suggest that, by considering only classical logic, Peirce sometimes advanced a misinterpretation of PSR°; he meant by the first word "Everything…", an unbounded number of determined facts or possible hypotheses and moreover he understood the change of the PSR° as a change from an infinite number of hypotheses to the best one, i.e. as a choice. In fact, this is the task that several statements by Peirce attributed to the abduction process.

which Peirce might hope to see formalised one day. According to the above interpretation, we can see in it a change in the kind of logic.

> Reasoning is a process in which the reasoner is conscious that a judgment, the conclusion, is determined by other judgments of judgements, the premises, according to a habit of thought, which he may be not able precisely to formulate, but which is approved as conducive to true knowledge. By true knowledge he means, though he is not usually able to analyse his meaning, the ultimate knowledge in which he hopes that belief may ultimately rest, undisturbed by doubt, in regard to the particular subject to which his conclusion relates. Without this logical approval, the process, although it may be analogous to reasoning in other respects, lacks the essence of reasoning (2.773).

Indeed, the following quotations show that Peirce's conception of abduction includes both logical reasoning and a non-logical step:

> [Abduction is a] conscious and controlled adoption of a belief [on the affirmative hypothesis concluding a PO theory] as a consequence of other knowledge [the previous argumentation of a PO theory on facts]" (2.442).
>
> On its side, the perceptive judgment [i.e. the affirmative hypothesis] is the result of a process, although of a process not sufficiently conscious to be controlled, or, to state it more truly, not controllable and therefore not fully conscious [i.e. the change of logic as an application of PRS°]. If we were to subject this subconscious process to logical analysis, we should find that it terminated in what that analysis would represent as an abductive inference [i.e. the change of logic], resting on the result of a similar process which a similar logical analysis would represent to be terminated by a similar abductive inference [the application of PSR°], and so on *ad infinitum*. This analysis would be precisely analogous to that which the sophism of Achilles and the Tortoise applies to the chase of the Tortoise by Achilles, and it would fail to represent the real process for the same reason. Namely, just as Achilles does not have to make the series of distinct endeavors which he is represented as making, so this process of forming the perceptual judgment, because it is subconscious and so not amenable to logical criticism, does not have to make separate acts of inference, but performs its act in one continuous process (5.181).

Indeed, since Goedel's result [59, 1933a] it is well-known that an infinite in some way divides classical from intuitionist logic, because there is no finite classical model of the latter.

Hoffmann adds:

> The astonishing fact is that, while logic necessarily seems to be *discrete*—self-controlled reasoning step by step, Peirce's logic of abduction depends on the continuity of an unconscious process [i.e. the application of PSR°],… my starting point is Peirce's claim that in abduction "the entire logical matter of a conclusion… must come from the uncontrolled part of the mind" (5.194) [10, p. 285].

That agrees with my view; Peirce implicitly referred to PRS°.

About this point Hoffmann stresses an important distinction:

> … Peirce uses "the distinction between the matter and the logical form" in order to show that "the entire logical matter of a conclusion must in any mode of inference be contained, piecemeal, in the premises. Ultimately therefore it must come from the uncontrolled part of the mind, because a series of controlled acts must have a first [in it]" (5.194). The essential point seems to be that, as a consequence of the *excluded* possibility that the "new" elements "first emerge in the conclusion of an[y] inference [of a deductive kind]", it must be assumed that they are "given" in some way "in a perceptive judgement (ibid.)

> The perceptive judgement is the ground of the *premises* [of the entire logical chain] and not of the conclusion… In this way the *logic* of abduction would have nothing to do with forming, creating or adopting hypotheses, but only with the *form* of inference mentioned above [i.e. the change from the classical logic to the classical]. And this "logical form" of the abductive inference enters the stage not before the original creative *act* is completed - in a perceptive judgement: "The first emergence of this new element into consciousness [as an affirmative hypothesis] must to be regarded as a perceptive judgment. We are irresistibly led to judge that we are conscious of it. But the connection of perception with other [subsequent] elements must be an ordinary logical inference, subject to error like all inferences" (5.192)…

> The explaining idea emerges in *perceiving* facts and experiences, and *not* in the conclusion of an inference… [this fact] can be explained by distinguishing between the logical or inferential form of [previous similar] abductive arguments and the genuine *process* of "getting" an explanatory hypothesis. There is a form of inference, but there are no rules [belonging to any particular logic] for getting a hypothesis.
>
> Peirce himself distinguished the *logical* or *inferential* side of abduction from its creative side by distinguishing self-controlled [i.e. deductive] "reasoning from the processes by which perceptual judgements are formed" and by claiming that"self-control of any kind is purely inhibitory. It originates nothing"(5.194). For that, I think, it is necessary to make a terminological distinction between *inferential aspects of abduction* and *perceptive aspects of abduction.* While the inferential side is characterised by the whole syllogistic formula mentioned in the above, the perceptive side, i.e. the genesis of the perceptual judgement, in which a hypothesis of the form "if A were true, C would be a matter of course" firstly emerges, has to be located at the premises. In order to explain within Peirce's concept of "logic" abduction as "the process of forming an explanatory hypothesis", *both* sides must come together"[10, pp. 278–280].

> … [In this way] we obtain, with Peirce, the rather paradoxical conception of a "logical inference" that is conceivable without logical rules [i.e. rules pertaining to any particular logic] (cfr. Peirce 5.188 and 7.220) [10, p. 278].

> … [In this way] we obtain, with Peirce, the rather paradoxical conception of a "logical inference" that is conceivable without logical rules [i.e. rules pertaining to any particular logic] (cfr. Peirce 5.188 and 7.220) [10, p. 278].

Hence, Hoffmann suggested:

> "… it is clear that the problem of an adequate description [of abduction] goes beyond the scope of logic and mathematical techniques. It is an *epistemological problem*, a problem of our knowledge about the world to find an adequate representation [10, p. 275].

Also Hintikka puts the problem to this level when he suggested that "… the validity of an abductive inference is to be judged by strategic principles, rather than by definitory (move-by-move) rules" [9, p. 513].

This agrees with Peirce's epistemological views:

> Retroduction goes upon the *hope* that there is sufficient affinity between the reasoner's mind and nature's to render guessing not altogether hopeless…. [Just what PRS° presupposes, as we saw in the above]

> It is somehow more than a mere figure of speech to say that nature fecundates the mind of man with ideas which, when those ideas grow up, they resemble their father, Nature (5.591).
>
> [The abduction is the] only possible hope of regulating our future conduct rationally (2.270).

These ideas amount to the premises of PSR°, i.e. the *ad absurdum* argument: "otherwise Nature would be irrational". Hintikka also reiterates the idea in the following terms: "It seems that our abductive hypothesis-forming power is nothing but a mysterious power of guessing right" [9, p. 505).

Moreover, at least on one important occasion Peirce's reasoning was very close to PSR°. It occurred when he illustrated the logic of his more general theory, the philosophy of pragmatism, as essentially based on abduction. He wrote:

> Admitting, then that the question of Pragmatism is the question of Abduction… What is the end of an explanatory hypothesis?… Any hypothesis, therefore, may be admissible, in the absence of any special reasons to the contrary, provided it be capable of experimental verification, and only insofar as it is capable of such verification (5.197).

Notice that the above DNS is regulating Peirce's entire idea; it expresses the formal predicate of PSR°: $\neg\exists x\neg f(x)$.

Result no. 5: *There exists an interpretation of Peirce's writings on abduction as closely approaching the logical scheme of reasoning of a PO theory which obtains a testable hypothesis according to PSR°.*

Let us recall Markov's principle; it is similar but different from what was previously presented as Peirce's principle:

$$\neg\neg\exists xf(x) = \; > \exists xf(x).$$

From Dummett's table [23, p. 29] we see that it is not equivalent to both the previous principle and PSR°; it is weaker than both (apart from the clauses of the Markov's principle, i.e. the decidability of the predicate and its occurring after an *ad absurdum* proof); indeed, while the conclusion of PSR° refers to the universe of discourse, Peirce's change obtains only an existential predicate. Peirce himself did not claim that his principle changes, as PSR° does, the entire logic of the theory into the classical one, which governs certainties; Peirce always considered the result to be a temporary certainty, a guess whose validity or non-validity is not known.

10 Conclusions

All the above offers a solution of the problem Black stated at the end of his paper on induction: "to elaborate a detailed and comprehensive account of scientific practice that will be reasonably close to the best actual procedures used in reasoning about matters of fact" [4, p. 179].

Black suggested also the following three basic problems concerning induction (in the sense of all non-deductive processes of inference):

> (1) The general justification problem. Why, if at all, is it reasonable to accept the conclusions of certain inductive arguments as true—or at least probably true?….
> (2) The comparative problem: Why is one inductive conclusion preferable to another as better supported?….
> (3) The analytical problem: What is it that renders some inductive arguments rationally acceptable? (Black p. 170)

According to the interpretation illustrated above, the first problem is solved since the above interpretation of an abduction relies on a logic, but non-classical, rather than classical logic. The second one is solved by the use of the *ad absurdum* proofs, which select and improve the hypotheses. The third one by means of the PSR°, which attributes a noble philosophical fatherhood to the inferential process.

Retrospectively, one may suspect that the greatest difficulty a scholar meets when inquiring into induction and abduction is precisely the widespread dogma which denied that DNSs are important, although it is well known that classical logic is absolutely unable to represent both induction and abduction; indeed, any inquiry into abduction, when confined to classical logic, results in no more than informal, or illative, extensions of deductive reasoning.

In the past the use of non-classical logic presented by several scientific theories passed unnoticed owing to several historical facts. First of all, in mathematical logic non-classical logic was not recognized as a respectable logic before the '30s of last century, when it was proved that even the logic of the prestigious quantum mechanics was non-classical. However, even after this, non-classical logic was ignored in the analyses of the theories of natural sciences, because quantum logic has still at present to receive a common definition.

Let us remark that from the previous analysis we have a definition of abduction:

"Abduction is a logical process aimed at solving a universal problem by the invention of a new scientific method, developed by reasoning through DNSs grouped into *ad absurdum* proofs, whose chain achieves a universal predicate on all cases implied by the given problem; the corresponding affirmative predicate plays the role of a hypothesis from which to deduce all possible testable consequences."

However, there exist some stringent conditions for applying such an abductive process:

(1) to meet a universal problem, unsolvable by current scientific means; (2) to reason by means of DNSs; (3) to prove by means of *ad absurdum* proofs; (4) to achieve a universal predicate regarding the given problem.

The above interpretation of the process of abduction is clearly a creative generation of new hypotheses, inasmuch as they are the beginnings of a new theory. This process proves to be demonstrative since indeed it reasons by means of *ad absurdum* proofs; it is ampliative since it enlarges our previous theoretical knowledge; it is almost additive since it adds knowledge which is of a plausible nature only (“imaginary”, as Lobachevsky called it).

All the above proves Peirce’s statement: “All the ideas of science come to it by the way of abduction” (5.145) as well McMullin’s thesis: Abduction is “the inference that makes science” [3] although the science of the theories of a particular kind, i.e. the PO theories.

All the above also suggests a solution to that problem which Hintikka stated to be the fundamental problem of contemporary epistemology, i.e. to clarify the notion of abduction [9].

Acknowledgments I am grateful to David Braithwaite for having revised my poor English and to an anonymous referee for some important suggestions.

References

1. Reichenbach, H.: Experience and prediction: an analysis of the foundations and the structure of knowledge. University of Chicago Press, Chicago (1938)
2. Peirce, C.S.: Collected Paper by Charles S. Peirce, vol. 1–8, pp. 1931–1958, Harvard, Cambridge (1931)
3. McMullin, E.: The Inference that Makes Science. Marquette U.P, Milwaukee (1992)
4. Black, M.: Induction, Encyclopedia of Philosophy. Mac Millan, London (1966)
5. Feferman, S.: In the Light of Logic, pp. 77–93. Oxford U.P, Oxford (1998)
6. Niiniluoto, I.: Defending abduction. Philos. Sci. **66**, 346–351 (1999)
7. Aliseda, A.: Abductive Reasoning: Logical Investigations into Discovery and Explanation, Synthese Library vol. 330. Springer, Berlin (2006)
8. Cellucci, C.: The scope of logic: deduction, abduction, analogy. Theoria **64**, 217–241 (1998)
9. Hintikka, J.: What is abduction? The fundamental problem of contemporary epistemology. Trans. C.S. Peirce Soc. **34**(503–533), 510 (1988)
10. Hoffmann, M.: Problems with Peirce’s concept of abduction. Found. Sci. **4**, 271–305 (1999)
11. van Heijenoorth, J.: Logic as Calculus and Logic as Language. In: Cohen, R.S., Wartowsfky, M.W. (eds.) Boston Studies of Philosophy of Science, vol. 3, pp. 440–446, Reidel (1963)
12. Lorenz, K.: Rules versus Theorems. J. Phil. Logic **2**, 352–359 (1973)
13. Hintikka, J.: The place of C.S. Peirce in the history of logical theory. In: Brunner, J., Forster, P. (eds.) The Rule of Reason. The Philosophy of C.S. Peirce, pp. 13–33. University of Toronto Press, Toronto (1997)
14. Brouwer, L.E.J.: The Unreliability of the Logical Principles, Collected Works 1908C. North-Holland, Amsterdam (1975)
15. Lewis, C.I., Langford, C.H.: Symbolic Logic. Dover, New York (1959)
16. Psillos, S.: An explorer upon untrodden round: Peirce on abduction. In: Gabbay, D.M., Hartman, S., Wood, J. (eds.) Handbook of History Logic, vol. 10, pp. 117–151. Elsevier, New York (2001)
17. Horn, L.: The Logic of Logical Double Negation. In: Proceedings of the Sophia Symposium on Negation, pp. 79–112. University of Sophia, Tokyo (2001)
18. Horn, L.: On the Contrary. In: Proceedings of the Conference Logic Now and Then, Brussels (2008) (in press)
19. Klein, M.J.: Some turns of phrase in Einstein’s early papers. In: Shimony, A. (ed.) Physics as Natural Philosophy, pp. 364–373. MIT Press, Cambridge (1982)
20. Drago, A.: History of the Relationships Chemistry-Mathematics. Fresen. J. Anal. Chem. **337**, 220–224 (1990)

21. Drago, A.: La maniera di ragionare di Lavoisier, Dalton ed Avogadro durante la nascita della teoria chimica", Rend. Acc. Naz. delle Scienze detta dei XL, Memorie di Sci. Fisiche e Nat., ser. V, 31, pt. II, tomo II, 189–201 (2007)
22. Lavoisier, A.-L.: (1862–1892). Oeuvres de Lavoisier, Paris, t. 1
23. Dummett, M.: Principles of Intuitionism. Clarendon Press, Oxford (1977)
24. Prawitz, D., Melmnaas, P.-E.: A survey of some connections between classical intuitionistic and minimal logic. In: Schmidt, H.A., Schütte, K., Thiele, H.-J. (eds.) Contributions to Mathematical Logic, pp. 215–229. North-Holland, Amsterdam (1968)
25. Grzegorczyk, A.: Philosophical plausible formal interpretation of intuitionist logic. Indag. Math. **26**, 596–601 (1964)
26 Troelstra, A.van Dalen, D.: Constructivism in Mathematics. North-Holland, Amsterdam, (1988) **1**.
27. Nickles, T. (ed.): Scientific Discovery, Logic and Rationality. Reidel, Boston (1980)
28. Achinstein, P.: Scientific Discovery and Maxwell's Kinetic Theory. Philos. Sci. **54**(3), 409–434 (1987)
29. Burch, R.: Charles Saunders Pierce. In Zalta, E.N. (ed.): Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/peirce-benjamin/ (2012)
30. Hilpinen, R.: Peirce's Logic. In: Gabbay, D.M., Woods, J. (eds.) Handbook of History of Logic, vol. 3, pp. 611–658. Elsevier, New York (2004)
31. Douven, J.: Abduction. In: Zalta, E.N.: (ed.) Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/abduction/ (2011)
32. Chiasson, P.: Abduction as an aspect of retroduction. Digital Encyclopedy of Charles S, Peirce. http://www.digitalpeirce.feeunicamp.br/p-abachi.htm (2001)
33. McKaughan, J.: From ugly duckling to Swan: C.S. Peirce, abduction, and the Pursuit Of Scientific Theories. Trans. C.S Peirce Soc. **44**, 446–468 (2008)
34. Stalker, D. (ed.) Grue, The New Ridde of Induction, Open Court, Chicago and La Salle,1994 (1990)
35. Flach, P.A., Kakas, A.: Abductive and inductive reasoning: background and issues. In: Flach, P.A., Kakas, A. (eds.) Induction and Abduction. Essays on their Relations and Integration, pp. 1–27. Kluwer Academic Publisher, Boston (2000)
36. Hanson, R.N.: Pattern of Discovery. Cambridge U.P, Cambridge (1958)
37. Fann, K.T.: Peirce's Theory of Abduction. Nijhoff, The Hague (1970)
38. Josephson, J.R., Josephson, S.G.: Abductive Inference: Computation, Philosophy and Technology. Cambridge U.P, Cambridge (1994)
39. Magnani, L.: Abduction, Reason, and Science. Process of Discovery and Explanation. Kluwer Academic Publisher, Dordrecht (2001)
40. Bazhanov, V., Drago, A.: A logical analysis of Lobachevsky's geometrical theory. Atti Fond. G. Ronchi, **64**(4), 453–481 (Italian reduced version A. Drago and A. Perno (2004), "La teoria geometrica delle parallele impostata coerentemente su un problema (I)", Periodico di Matem., ser. VIII, **4**, ott.-dic., 41–52) (2010)
41. Drago, A.: A.N. Kolmogoroff and the relevance of the double negation law in science. In: Sica, G. (ed.) Essays on the Foundations of Mathematics and Logic, pp. 57–81. Polimetrica, Milano (2005)
42. Drago, A.: Storia del corpo nero: la strategia effettiva di Planck rivelata dalle sue scelte fondamentali. Atti Fond. Giorgio Ronchi. **5**(65), 619–634 (2010)
43. Drago, A.: La teoria delle relatività di Einstein del 1905 esaminata secondo il modello di organizzazione basata su un problema. In: Giannetto, E., Giannini, G., Toscano, M. (eds.) Relatività, Quanti, Caos e altre rivoluzioni della Fisica, pp. 215–224. Guaraldi, Rimini, (2010b)
44. Drago, A.: Pluralism in logic: the square of opposition, Leibniz' principle of sufficient reason and Markov's principle. In: Béziau, J.-Y., Jacquette, D. (eds.) Around and Beyond the Square of Opposition, pp. 175–189. Birkhaueser, Basel (2012)

45. Drago, A.: The emerging of two options from Einstein's first paper on quanta. In: Capecchi, D., Pisano, R. (eds.) Proceedings of the XXXII Congress SISFA 2012. Springer, Berlin (in print) (201?)
46. Drago, A., Oliva, R.: Atomism and the reasoning by non-classical logic. HYLE **5**, 43–55 (1999)
47. Drago, A., Pisano, R.: Interpretation and reconstruction of Sadi Carnot's Réflexions through non-classical logic. Atti Fond. Ronchi, **59**, 615–644 (Italian version in Giornale di Fisica, **41** (2000), 195–215) (2004)
48. Einstein, A.: Zur Elektrodynamik bewegter Körper. Annalen der Physik **17**(891–921), 891 (1905a)
49. Markov, A.: On Constructive Mathematics. Trudy Math. Inst. Steklov, **67,** 8–14; Engl. tr. in Am. Math. Soc. Translations, (1971) **98** (2), 1–9) (1962)
50. Carnot, S.: Réflexions sur la puissance motrice du feu. Blanchard, Paris (1824)
51. Lobachevsky, N.I.: Geometrische Untersuchungen zur der Theorien der Parallellineen, Finkl, Berlin (English transl. as an Appendix to R. Bonola (1955), Non-Euclidean Geometry. Dover, New York) (1840)
52. Kolmogorov, A.N.: On the principle 'tertium non datur'. Mathematicheskii Sbornik, **32**, 646–667 (Engl. transl. in J. van Heijenoorth (1967), From Frege to Goedel. Harvard U.P., Cambridge, 416–437) (1924/25)
53. Avogadro, A.: Essay d'une manière de detérminer les masses relatives des molécules élémentaires des corps…. Journal de physique, de chimie et d'historie naturelle **73**, 58–76 (1811)
54. Kleene, C.: Introduction to Metamathematics. Van Nostrand, Princeton (1952)
55. Einstein, A.: Ueber einen die Erzeugung der Verwandlung des Lichtes betreffenden heuristisch Gesichtpunkt. Annalen der Physik. **17**, 132–148 (Engl. Tr. Am. J. Phys., 33 (1965) 367–374) (1905b)
56. Tennant, N.: Natural Logic. Edinburgh University Press, Edinburgh (1990)
57. Achinstein, P.: The Book of Evidence. Oxford U.P, Oxford (2001)
58. Fisch, M.H.: Peirce and Leibniz. J. Hist. Ideas **33**, 485–496 (1972)
59. Goedel, K.: Collected Works, vol. I. Oxford U.P, Oxford (1986)

Freedom and Moral Judgment
A Cognitive Model of Permissibility

Sara Dellantonio and Luigi Pastore

Abstract Contemporary research in the fields of moral psychology and cognitive philosophy has provided considerable data supporting the claim that there are important similarities in the ways in which different people conceive of morality and produce moral judgments. However, one of the more pressing questions is how to account for the fact that, despite these similarities, moral judgments appear to be highly variable both on a cultural and individual level. This paper addresses this issue by developing a model which is inverted with respect to the one usually embraced by the cognitive literature on morality. Instead of analyzing the problem of moral judgment starting from all the actions that are considered impermissible, this work assumes that people first judge which actions are morally permissible. Permissibility is interpreted in terms of what each subject feels he/she must be free to do. The advantage of this inversion is that it allows us to make a connection between two lines of research that are usually considered unrelated: research on the processes underlying the production of moral judgments and research on the problem of determining how people understand 'freedom'. Regarding this latter issue, the article focusses specifically on George Lakoff's cognitive analysis of how humans develop their concepts of freedom. The starting point of Lakoff's analysis is that different groups and different individuals do not have the same understanding of 'freedom', even though everybody shares the common empirical core concept. Lakoff puts forward a model which aims to explain both the common cognitive ground of the various concepts of freedom and how these concepts vary depending on other cognitive elements connected to them. In this work we try to show that Lakoff's model can provide an explanation of moral judgment that

S. Dellantonio (✉)
Università degli Studi di Trento, Trento, Italy
e-mail: sara.dellantonio@unitn.it

L. Pastore
Università degli Studi di Bari Aldo Moro, Bari, Italy
e-mail: luigi.pastore@uniba.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_19, © Springer-Verlag Berlin Heidelberg 2014

accounts for both the cross-cultural and trans-individual similarities and the cultural, individual and situational differences.

1 Introduction

Contemporary debate in the fields of moral psychology and cognitive philosophy addresses the question of moral judgment starting from the idea that human beings are characterized by specific capacities that allow them to distinguish morally permissible, impermissible, and obligatory actions. As summarized by Susan Dwyer: "A […] striking fact about our species is that all (normal) humans develop into moral agents, that is, into creatures with (at least) the following moral capacities: the ability to make judgments about the moral permissibility, moral impermissibility, and moral obligatoriness of actions in actual and hypothetical, novel and familiar cases; the ability to register morality's special authority (i.e. the fact that moral imperatives are nonhypothetically binding and sometimes contrary to self-interest); the ability to make attributions of moral responsibility for actions (as distinct from attribution of mere causal responsibility); and the ability to recognize the force of excuses." [12, pp. 237–238] To address morality from a cognitive point of view we need to investigate the psychological processes that lead people to develop their moral views and to produce moral judgments.

The contemporary cognitive debate tends towards a nativist view, whose main effort is to identify transcultural invariants and universal principles of moral judgment (for a review of this tendency see also e.g. [7, 10]. Even though cognitive research has provided many data supporting the claim that there are important similarities in the ways in which different people (small children and adults, as well as people belonging to different cultures: for reviews, see [31, 35, 39]) conceive of morality and produce moral judgments, we also need to face the problem of the continual variability of moral judgments. In fact, despite similarities, people belonging to different cultures or groups rely on moral principles which often differ widely (see e.g. [33]). Furthermore, people belonging to the same culture or group often embrace very different moral principles and also the moral judgment of the same person can change over time. Likewise, the application of moral principles is continually subject to exceptions and situational 'readaptations' so that they are applied and weighted differently in different cases. Indeed, it is possible that a person considers it in general morally impermissible to kill another human being, but that his/her moral judgment would reverse in a case where the person in question is a criminal or a terrorist. Or it is possible that someone strongly believes that it is morally obligatory to help a human being whose life is at risk, but this does not exclude that his/her judgment can change in a case where the person in question is e.g. an illegal immigrant arriving on a boat. Because of this intrinsic ambiguity of moral judgment which is in some aspects highly stable and in others highly variable, we need to develop a model that accounts not only for cross-cultural and trans-individual similarities but also for the variability of moral judgment with respect to different cultures or groups, different individuals, and the same person over time and in different situations. This paper tries to sketch such a model starting from the theory George Lakoff develops on how people understand freedom and on what people think they must be free to do (i.e. about what people think it must be permissible).

Usually the issue of moral judgment is addressed by first analyzing all the actions that are considered *impermissible* and the rules that *prohibit* certain behaviors, more rarely it is addressed by first analyzing the actions that are considered *obligatory*, while permissible actions are seen as a logical residual with respect to these: an action is permissible when it is neither impermissible, nor obligatory. The reason why the problem of moral judgments is generally discussed starting from impermissible or obligatory and not from permissible actions is that the field of permissible actions consists of all kinds of actions—including actions that have nothing to do with morality. Indeed, the presence of a possibly morally relevant situation is indicated by the fact that a certain action is considered impermissible or obligatory. However, the fact that impermissible and obligatory actions are indicators of possibly morally relevant situations does not necessarily imply that psychologically the field of impermissible actions is identified *before and independently* from the field of permissible actions and that moral judgments depend primarily on what we consider impermissible or obligatory and not on what we consider permissible.

This 'inverted model' interprets permissibility in terms of what each subject feels he/she must be free to do: permissibility describes the freedoms each subject perceives he/she must legitimately have. Impermissibility and obligation are derived from this idea of permissibility: the actions that overstep the limits of one individual's freedom and encroach on the freedom of someone else are perceived as impermissible, while the actions that protect the freedom of someone else against invasion are perceived as obligatory. This inversion allows us to link the debate on moral judgment with George Lakoff's cognitive analysis of the concept of freedom and how people understand freedom—i.e. on what people think it should always be permissible for them to do.

The starting point of Lakoff's analysis is that different groups and different people understand freedom differently. Lakoff tries to explain why different people have different concepts of freedom and puts forward a model to explain both the common cognitive ground of these concepts and the ways they vary depending on other cognitive elements connected with them.

The original contribution of this article lies firstly in pointing out that our idea of freedom is connected with our moral positions and that Lakoff's model of freedom can therefore be applied to an apparently different debate concerning moral judgment. Secondly, this work aims at showing that Lakoff's model gives important clues to answer some open questions in the debate on moral judgment and indirectly offers a highly flexible and strong explanatory model of moral judgment.

2 Cultural and Individual Differences in Moral Judgment

Despite all their specific differences, a large number of cognitive approaches to moral judgment share the general idea that the capacity to produce moral judgments is based on universal mechanisms that can be traced back to an innate, modular structure of the cognitive system.[1] Such a module should be able to select all and only the morally relevant actions and to analyze them identifying their consequences, who is affected by them and who carries them out as well as the intentions of these actors. Furthermore, this module should be set up with some sort of internalized moral principles or rules, whose function is to determine whether the considered actions are impermissible, permissible or obligatory (for a more detailed explanation see e.g. [19]).

The aspect most of the literature focuses on is the definition of the universal moral principles this moral module relies on. Some authors think that these principles consist of innate moral rules. Pinker and Brown maintain for example that moral capacities rest on universal rules consisting of specific prohibitions such as those against murder, harm and rape [32, p. 414, 5, pp. 138–139]. An attempt to precisely identify a possible universal rule of this kind has also been made by John Mikheil through his principle of the prohibition of intentional battery which suggests that people consider any illicit bodily contact that leads to physical harm as impermissible [27, 28, pp. 1057–1127].

Analogous conclusions concerning universally shared moral principles have also been drawn by authors who belong to a different research area whose primary aim does not consist of investigating moral capacities from the point of view of the organization of the mind into faculties or modules, and which focuses instead on the moral development of children with respect to their interaction with the environment (see e.g. [31, 40, 41], Smetana 2006, [37]). This kind of inquiry—which in the last decades has produced a substantial number of empirical studies in support of its general thesis (for a review see e.g. [31, 36])—starts from the idea that psychologically moral rules are perceived as different with respect to other kinds of rules such as conventional ones (like: don't eat with your hands; don't go to work wearing pajamas, etc.). Indeed, these studies suggest that moral rules are perceived as absolutely objective and valid in any context or culture and that their prescriptive force is considered independent of any authority, while their violation is believed to be particularly serious. These characteristics strongly distinguish moral from conventional rules which are perceived in contrast as less serious, more arbitrary and dependent on a specific (historical, cultural, institutional …)

[1] The existence of an innate moral faculty is hypothesized by e.g. [6], pp. 152–153, [18, 19, 27, 28, 30]; Haidt (see e.g. [13]) also speaks of a moral module, even though he maintains a different view according to which this module works with principles that are entirely derived by the culture the individuals belong to.

context [31, 35, 40, 42]. Empirical studies show that both toddlers and adults of different backgrounds make this kind of distinction between moral and conventional rules (for reviews see e.g. [31, 35, 39]). As far as the content of these rules is concerned, what distinguishes moral from conventional rules is that moral rules apply to different kinds of situations involving harm, justice, and rights, while conventional rules are related to other aspects of social life. “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” [42, pp. 169–170]. Experiments investigating the moral/conventional distinction generally make use of *prototypical* conventional and moral rules. As e.g. Kelly and colleagues point out: “moral rules typically involve a victim who has been harmed, whose rights have been violated, or who has been subject to an injustice”, while prototypical conventional rules do not involve issues relating to harm, justice, and rights [21, p. 118]. These definitions and these examples of moral rules are implicitly based on the assumption that morality relies on universal moral principles which consist mainly of a “general prohibition against harm, or at least against harming the innocent” [34, p. 373] as well as a prohibition against unjust behaviors and an obligation to respect the rights of others.

This general tendency of contemporary moral psychology to explain human moral capacities by making appeal to universal principles or rules of psychological origin tends however to eclipse the evident differences in the moral judgments produced by different cultures, groups or individuals. Even the most universalist views must indeed take into account a certain degree of variability in moral rules (see e.g. [19]. This is mostly considered in terms of cultural differences while the explanatory strategy that is usually used to conciliate universalism with cultural variability is to hypothesize that the principles the innate moral mechanisms work with are developed through contact with the moral rules embraced by the culture or group the subject grows up in and might therefore vary.

One of the best known authors who promotes a view of this kind is Jonathan Haidt. Haidt’s studies, which have been carried out in cultures other than the ones usually considered in academic research (i.e. mainly the North American and European ones) or include conservatives from Western cultures, show that people may consider as properly moral (and not just as conventional) not only issues relating to harm, rights and justice[2]—as suggested e.g. by Turiel and colleagues—but also issues of an entirely different kind relating to in-group/loyalty, authority/respect and purity/sanctity (see e.g. [15–17]). As Haidt points out: “in most cultures the social order is a moral order, and rules about clothing, gender roles, food,

[2] Haidt defines them as issues relating to harm/care and fairness/reciprocity.

and forms of address are profoundly moral issues. [...] In many cultures the social order is a sacred order as well." [17, p. 371] He suggests that the moral model works with specific principles drawn from the culture the judging subject belongs to (see e.g. [13, 14]).[3]

Even though Haidt assigns the leading role in determining the moral judgments of group members to cultural factors, his hypothesis does not offer any answer for the problem presented above of the (cultural and individual) variability of people's moral views. On the one hand it does not account for the wide individual (infra-cultural) variations of moral judgments.[4] On the other hand, it does not explain the connection between a moral system and specific cultural beliefs; e.g. it does not explain why *only some* groups or cultures consider issues related to in-group/ loyalty, authority/respect and purity/sanctity as also specifically moral, while issues related with harm, with the presence of a victim, with justice and rights are considered specifically moral by all groups or cultures (about this see [9]).

More generally, all of the various attempts to provide a model of moral judgment based on the idea that it is produced by a dedicated module (more or less influenced by cultural factors) face a common problem because they result in a description of moral judgment which is too rigid. In fact, moral judgments are always subject to great individual and situational variability which cannot be explained on the basis of inflexible mechanisms. If we take into account any set of moral principles or rules that state which actions are permissible, impermissible or obligatory in a specific group or in general, we will always find many cases in which they are partially subject to exceptions or even radically violated. The number of exceptions and violations will be so great as to put in doubt the very existence of such rules.

Part of this problem is discussed e.g. by Susan Dwyer who shares the idea that morality can be explained in terms of a capacity realized through a modular faculty, which she describes by analogy with Chomsky's language faculty (about this see [8]). According to Dwyer, the agent's permissibility/impermissibility judgments can be treated as analogous to the acceptability/unacceptability judgments speakers express with respect to sentences they consider grammatically adequate, while Chomsky's notion of a parameter constitutes a possible solution for the problem of the intrinsic variability of moral judgments in different contexts. Chomsky's notion of a parameter fulfills the function of explaining the fact that—even though generative grammar relies on universal principles, shared by all real

[3] For a critical examination of Haidt's view showing—among other things—that the moral principles identified by Turiel and colleagues have a greater degree of universality compared to the ones identified by Haidt: see [9].

[4] Haidt's view tries to respond to the need for explaining why different cultures rely on different moral principles. However, *de facto* it binds moral evaluation entirely to cultural rules concluding that a virtuous person is a fully enculturated person ([14, p. 216]). Thus, Haidt bars the possibility of explaining those cases in which an individual moral judgment diverges from the rules expressed by the culture he/she belongs to and therefore fails to account for the individual and situational variability of moral judgment (on this issue see also [9]).

and possible languages—some basic features of the grammar of natural languages are subject to variations that are framed around a limited and defined range of possible options. An example of principles and parameters in natural languages can be the one of grammatical subject: no language can work without grammatical subject (principle); however languages can be divided in two groups: the ones where the subject must be explicitly expressed and the ones where it need not be (parameter). The presence of this parameter allows—on the basis of minimal linguistic experience—to set linguistic learning to the right options, corresponding to the rules of the specific language the child is exposed to, so that the child can learn the rules of this language.

As for moral judgment, according to Dwyer the notion of parameter is useful in explaining how—given any set of universal moral principles—the application of moral judgments in different groups or cultures is subject to variations. For Dwyer these variations are not found in the moral principles themselves, but rather in the kinds of 'entities' they apply to: "One thing seems to be true of all known human moral systems: moral considerations (obligations and prohibitions) do not apply to everything. For example, pieces of furniture are not the sorts of thing that have moral considerability [...]. Some human moral systems cast the net widely, including all animals along with human beings; other are more conservative, extending moral considerability only to human beings (and perhaps only to a subset of human beings—what moral philosophers like to call *persons*). [...] some such systems might assign different degrees of moral considerability to different types of members of the class, ranking say, human beings above nonhuman animals, men above women, or cows above frogs." [12, p. 249].

To explain and to bind these variations to specific criteria Dwyer introduces the notion of "schweeb" which is interpreted as a parameter of Chomsky's kind. "Let us [...] define a schweeb as a 'creature with the highest moral status'. A very basic principle of all possible [internalized] moralities might be 'Schweebs are to be respected' or 'Given the choice of saving the life of a schweeb or saving the life of a non-schweeb, always save the life of a schweeb.'"[12, p. 249] According to Dwyer, "schweeb"/"non-schweeb" is a parameter that—analogously with Chomsky's parameters—is set during learning, is drawn from the culture or groups to which the subject is exposed to in an early stage of his/her development and establishes which subjects deserve the maximum moral considerability and therefore moral protection and which do not.

'Always protect schweebs!' is a kind of moral imperative which is present, according to Dwyer, in the mind of each individual. An action that appears to be absolutely morally impermissible if carried out at the expense of a schweeb might appear morally permissible or even obligatory when it is carried out at the expense of a non-schweeb. An example of absolute separation between "schweebs"/"non-schweebs" that fits with Dwyer's description of this notion could be societies that practice slavery, where slaves were excluded from having any moral status at all, or in situations characterized by strong racial prejudices or conflicts where specific ethnic groups aren't accorded any binding moral status in virtue of which they need to be respected (think for instance of the Jews in Germany during the Nazi period or more generally of any kind of genocide that has happened in history). A slave or Jew wasn't accorded any moral protection with respect not only to their individual rights, but also to their physical integrity and to their life. On the contrary, they suffered maximum restraints on all their fundamental rights, which resulted in the legitimatization of all forms of violation perpetrated against them. An action that is absolutely impermissible when it is directed at a free man—or, in a context like the Nazi one, against an Arian—becomes permissible or in some cases even obligatory when addressed against a slave or a Jew.

However, these extreme examples also immediately reveal the limits of a position that considers 'schweeb' as a parameter. A parameter should in fact be uniform within a group, should be set early in cognitive development, should not be subject to modifications over time or to massive exceptions or variation within the same group. On the contrary, history and everyday life suggest that moral judgments are highly variable both within groups and over time. If we consider for example the case of the Germans during the Nazi period, we see that individual positions about the degree of moral protection which should have been granted to Jews was neither uniform for all Arians nor invariable over time. Not all Arians agreed with the racial laws and with the idea that Jews shouldn't be granted moral consideration or protection (or that they should have been less morally considered or protected than Arians). Furthermore, the position of many Arians changed over time: even though some people initially embraced the Nazi ideology, later they reconsidered their position and in some cases even put their own life at risk to protect Jews from deportation. Moreover, not all Jews were considered equal with respect to the degree of moral protection assigned to them: intellectuals and the leisured classes received special treatment with respect to ordinary people because of their merits, their connections and their socio-economic status (see e.g. [1]).

Contemporary Western society has largely embraced, at least from a formal point of view, the position stated in the *Universal Declaration of Human Rights* that all human beings are equal and therefore—according to Dwyer's use of the notion of parameter—equally 'schweebs'. However, in spite of this, individuals' judgments about the degree of moral protection that should be granted to other human beings varies widely depending on the situation and on the people involved into it. All the laws against terrorism recently passed with the approval of public opinion show for example that individuals culpable of what are considered terrorist acts are generally not perceived as worthy of the same moral consideration and protection as ordinary people (for a socio-psychological analysis of the moral consequences of depicting someone as a terrorist see e.g. [2, 20, 26]. In fact, something similar happens in all cases where we face a serious violation of the moral norms we believe in. People usually accord criminals a lower degree of moral consideration. Even criminals themselves perceive people who committed crimes they consider particularly dreadful as less worthy of moral consideration and protection than other moral beings. Think for example of the fact that when child offenders enter the criminal justice system, they tend to be harassed even by other criminals.[5] A general principle that can be applied in all these cases is that *when a person is morally culpable of crimes* considered particularly serious, he/she is also perceived as deserving less moral protection (see also e.g. [4]). In this sense, the supposed schweeb-parameter is not applied in all situations in the same way, but is continually subject to exceptions depending on many other factors including the specific person we are considering.

Dwyer is aware that her model can be criticized on the basis of arguments of this kind related—among other things—to the variability of the individual's moral judgment. Indeed, as she points out: "Members of the same families, exposed to virtually identical environments, disagree about the permissibility of same-sex sex, abortion, and eating nonhuman animals." [12, p. 251] The solution Dwyer suggests to respond to these objections while preserving the analogy between moral faculty/moral judgment and generative grammar/grammaticality judgment is to abdicate the notion of (shared) Language and maintain instead that different individuals speak subjective idiolects: everyone follows implicitly different grammar rules that belong to his/her personal language and that lead him/her to give different judgments about which sentences are grammatically correct. Analogously, in the moral field, what people interiorize are individual moral rules—metaphorically we could say 'idiolectical moral rules'—which lead them to produce different judgments about which actions are morally permissible.

This solution presupposes that parameters are set differently by different subjects and explains on this basis individual variability in moral judgments within the same group or culture.[6] However, the problem of this view is that *de facto* it gives up the idea that it is possible to develop a general model of moral judgment: every individual is different, speaks his/her own personal moral idiolect and produces permissibility judgments according to the specific, unique rules of his individual moral grammar. In this sense, even though Dwyer has the merit of stressing that theories about moral judgment need to take into account both the high individual variability of moral judgment and the variability of the parameter that specifies who needs to be morally considered and protected and who does not, her solution consists ultimately in the abdication of the search for a general model of moral judgment. In sum, *it implies the impossibility of specifying the regularities beyond the variability* and of determining e.g. why a person embraces certain rules instead of others, why people sometimes refuse rules they previously embraced, what exactly changes in these cases, how different principles are related to each other, etc. If we agree that every individual follows his/her own moral rules which might differ from anyone else's and even change over time, and if we think that there

[5] M. S. James, Prison Is 'Living Hell' for Pedophiles, http://abcnews.go.com/US/story?id=90004#.T7o8BMXiYjk,abcNEWS Aug. 26, 2003.

[6] It is not clear how this solution can preserve the analogy with Chomsky's theory in relation to many factors, among others: what are the universal principles beyond this variability; why do people change their 'moral parameters' over time, when they should be set once and for all at an early developmental stage; how is it possible that 'moral parameters' are subject to exceptions, etc.

isn't any general model to explain this individual variation, we can only describe the singular moral judgments of a person at a given time, but we give up the possibility of explaining what determines them: i.e. why a person in a given time produces a certain judgment instead of another. To establish just such a general model is one of the main aims of cognitive approaches to morality.

3 Moral Permissibility and Lakoff's Cognitive Model of "Freedom"

Most of the theories of moral judgment implicitly identify the moral field on the basis of the actions that are considered impermissible or in some cases obligatory and on the basis of the rules stating which actions are prohibited or obligatory. As Shaun Nichols pointed out with regard to studies investigating the moral/conventional distinction: "in typical moral scenarios presented in the moral/conventional task, people's judgments are guided by an internally represented body of information, a 'normative theory', prohibiting behavior that harms others." [30]. The reason why most moral models start from actions considered impermissible or obligatory is quite intuitive: permissible actions include all kinds of actions, not only morally relevant ones; it is the fact that certain actions are considered impermissible or obligatory that univocally indicates their moral relevance.

However, the fact that it is the impermissible and maybe the obligatory that provide clues to identifying morally relevant actions and distinguishing them from non-morally relevant actions does not necessarily imply that from a psychological point of view—when we are faced with morally relevant actions or situations—we first produce a judgment of moral impermissibility (or obligation) and then determine which actions are permissible by exclusion, considering permissible all actions that are neither impermissible nor obligatory. On the contrary, it is possible that we first produce a permissibility judgment. In fact, in this part of the article we will sketch a model that starts from permissibility and deduces from it both impermissibility and obligation. The advantage of this so to speak, 'inverted model', is that it allows us to explain moral judgment in a more flexible way, integrating in a structured manner the cultural and the individual/situational variability of moral judgment.

In this inverted model permissibility is described in terms of what each individual perceives he/she can do, i.e. what each one feels he/she must be *free* to do since in a social context people are not free to do everything they wish, so everyone's freedom ends where the freedom of someone else begins. It is this reciprocal limitation on an individual's freedom that gives rise to impermissibility. A person will perceive as impermissible all the actions that go beyond the freedoms of one subject and invade the freedoms of another. This interpretation of impermissibility as derived from permissibility is strictly related to the idea that everyone strives for protecting his/her freedom. Obligation derives from the principle that it is necessary to protect freedom and therefore to coordinate the freedoms of different people when they tend to collide with each other.

This approach to the moral problem centered on permissibility allows us make a connection between the debate on the psychological factors behind moral judgments and George Lakoff's cognitive analysis of the concept of freedom and the way people conceive of freedom. By making this connection, we can *describe the way we produce moral judgments on the basis of the particularly effective and flexible model Lakoff has worked out to explain why different people conceive of freedom differently, what these differences depend on, why there is also something common among the different perspectives on freedom, and what this commonality consists of.*

Lakoff's starting point is perfectly coherent with the remarks proposed in the previous discussion regarding the wide cultural and individual variability of moral judgments. Different cultures and different people have different views on freedom (i.e. on what should be considered permissible). Analogously, according to Lakoff different cultures and different people have different views on what should be the specific extension and nature of individual freedoms. However, Lakoff thinks that the variation in individuals' concepts of freedom is neither complete nor casual, since everyone's concept of freedom rests on the same basic experience and is developed according to a common path.

Lakoff maintains that freedom is an "essentially contested concept" [23] with a fluid structure, whose exact content will always be the object of dispute because it can vary widely from one individual to another. Not only different people, groups and cultures have different concepts of freedom, but it is also clear that *the specific content of someone's concept of freedom depends on his/her political and ideological views*. One of the main examples discussed by Lakoff is the difference between the concept of freedom US conservatives and progressives rely on: "When a hard-core conservative uses the word "liberty" and applies it to "economic liberty" and "religious liberty", he has in mind the conservative version of the word's meaning. And when a progressive uses the same word—that the freedom to marry is a matter of a gay person's liberty—he has in mind a progressive version of the concept's meaning." [24, pp. 178–179] However, beyond the great differences that characterize concepts of freedom differently connoted from a political and cultural point of view, according to Lakoff the common intuition suggests not only that all societies and human groups have some concept of freedom—even though possibly only embryonic or different from the ones spread by contemporary Western culture—but also that all different concepts of freedom embraced by different people or groups share an "uncontested core" of "central meanings that almost everyone agrees on" [23]; see also [23] and [24, pp. 178–180].

The presence of a common core is due to the fact that, "though it is abstract, [the concept of freedom] is grounded viscerally in bodily experience." [23]. According to Lakoff humans develop their concepts of freedom starting from a specific perceptual experience that happens quite early, i.e. the experience of being free to move their own body: "The most basic idea of freedom is freedom of motion" [24, p. 180], see also [23]. Small children form their own intuitive concept of freedom starting from the (natural and fulfilling) experience of being able to move their body as opposed to the experience of not being able to move as they wish or of being kept from moving, which is viscerally perceived as oppressive, overwhelming, frustrating and even painful. This concrete concept of freedom as freedom to move is the cognitive basis for the development of our adult, abstract concept of freedom, which is built through a gradual extension of the core.[7] However, this experiential core is too concrete to determine 'freedom' in complex social and relational contexts. Since it is based on bodily experience, it doesn't say anything about the political and the social (i.e. about the abstract) meaning of "freedom". In fact, the adult, fully-developed, abstract concept of 'freedom' we use to evaluate complex situations is much broader then the core concept and it includes much more information. This is formed through an extension of the core which allows us to specify all the aspects that are needed in social situations which are mainly concerned with the problem of determining where the freedom of someone ends and where the freedom of someone else begins.

Lakoff suggests that the extension of the core concept to the fully developed abstract concept of freedom we use to weigh up social and political situations is accomplished by including the concept of freedom in *models*—called *Idealized Cognitive Models* (*ICMs*)—which are structured mental theories through which people interpret a complex piece of reality.[8] To explain the function of this structure in determining concepts, Lakoff considers, among others, the example of 'Tuesday'. Even a simple concept like 'Tuesday' cannot be determined and learned individually. To understand and to master it, people need to develop something like a mental theory—a *model*, according to Lakoff's terminology—which describes a specific general view of time and its organization into a linear sequence composed by parts defined by the movement of the sun, called days, and by a larger seven-day calendric cycle [22, p. 68]. An ICM can also be very complex and often depends on forms of information organization produced by specific groups or cultures. For example the organization of days in weeks as a

[7] Immediately after the core concept is formed, we naturally, automatically and unconsciously extend it through metaphorical thought which accomplishes the function of linking "abstract ideas to visceral bodily experience." [23]. The first extension is obtained by applying a so called "primary metaphor" which is basically universal and gives rise to a larger common core of the concept. (About primary metaphors (see e.g. [25, p. 46.]) The last and most important extension of the concept, as a consequence of which it takes its specific and individually different final abstract form, is carried out by inserting it into a so-called *cognitive model*: this step will be discussed below. For a more detailed analysis of Lakoff's theory on the origin and development of the first stages of the concept of freedom see also [11].

[8] The word cognitive indicates that they are mental theories: i.e. models internalized by the subject to understand the structure and the interconnections of specific complex phenomena, and to create links between concepts that allow their reciprocal determination. The word idealized suggests that the structures these models consist of "do not exist objectively in nature", but "are created by human beings" [22, p. 69].

whole with seven parts is not produced by individual minds; rather we inherit it from the way our society and our culture organizes time. In this sense, ICMs are not fully individual, since they are—at least in part—drawn from the social context people come into contact with. However each person might internalize social models differently: so, the ICMs of different people might also differ widely.

As far as 'freedom' is concerned, the ICM is the means by which the core is enlarged to obtain its final, complex form as a political abstract concept. Through its integration in a ICM the concept of freedom is immediately related to other concepts belonging to the same system like 'justice', 'equality', 'human nature', 'competition', 'property', 'right' etc. ([23] ff.) so that *these concepts mutually determine and specify each other in a quasi-holistic manner*. Only in this enlarged form can 'freedom' serve as a means for the subject to interpret and to weigh up the social reality around him. Even though individuals' ICMs are largely derived from their social context and from the social and political views that, implicitly or explicitly, characterize them, each person develops a specific ICM which never completely conforms to those developed by others. An ICM describes the specific view of freedom each individual develops and in turn gives rise to his/her specific understanding of the concept, together with his/her way of using it in thinking: "Each person has a concept of freedom that makes sense to him or her. […] For that person, her concept of freedom is the concept of freedom. She uses it to think with." [24, p. 178].

The ICM each of us develops/internalizes to understand a real context relies largely on what Lakoff calls 'deep frames', which are "mental structures of limited scope, with a systematic internal organization. For example, our simple frame for 'war' includes semantic roles: the countries at war, their leaders, their armies, with soldiers and commanders, weapons, attacks and battlefields. The frame includes specific knowledge: In the United States, the president is the commander in chief and has war powers; the purpose of war is to protect the country; the war is over and won when the other army surrenders. All words are defined with respect to frames." ([23, p.10–11], see also [23, p. 112] sgg.) Therefore, a frame is meant as something like a definition or, more precisely, a defining structure connecting one concept with some knowledge that specifies its content. As Lakoff describes it, frames are the molecules ICMs are composed of: their structure is analogous to that of an ICM, but they are less extended and more precisely defined.

When the original, empirical core concept build on the experience of being free to move grows through its inclusion in a ICM composed of frames to constitute the abstract concept of freedom we use in everyday life, the intrinsic ambiguities and blanks of the core concept with respect to its social and political dimensions are filled in. The content of the concept is specified by means of the definitions provided by the frame and by the general structure of the ICM. Because different ICMs are structured in a (slightly or widely) different way and are composed (in small or large part) of different frames, when the core concept is integrated in different ICMs it assumes different forms. Different ICMs will therefore give rise to concepts of freedom which are characterized by different political and ideological orientations. So, for example: "When the blanks are filled in by progressives and conservatives, what results are two radically different ideas expressed by the same word, 'freedom'." [23, p. 15].

However, the fact that individuals' ICMs and frames are different does not mean that their internal structure and their interconnections cannot be described and understood. But, if we can identify and describe the specific ICM in which an individual concept of freedom is integrated, we can also give a model of that person's concept of freedom and we can understand how he/she thinks with respect to freedom as well as with respect to all other issues related to freedom. It is this possibility of capturing the ICMs and frames which determine our individual concepts of freedom that can help us to clarify the problem of moral judgment and to develop a model built on the idea that permissibility describes the freedoms each subject perceives he/she must legitimately have.

The concept of freedom and therefore our idea of what should be permissible are—as we said before—both highly individual and dependent on social and cultural influences. However, their variation is not casual or inexplicable, but is due to the specific ICM the concept of freedom is integrated into. Since the ICM by which our concept of freedom is determined is constituted also by other crucial concepts such as 'justice', 'equality', 'human nature', 'competition', 'property', 'right' etc. and these concepts all mutually determine each other, the concept of freedom a person has will depend on how this person conceives of other things like justice, equality, human nature etc. Lakoff's suggestion is that the constituents and reciprocal interconnections of an individual's ICM can be identified and described thereby helping us determine what exactly freedom consists of for this person. This description would allow us to determine *when* (in what cases) and *why* (for what reasons) a particular person will perceive a specific action carried out in a specific context as permissible (i.e. as free) rather than as impermissible or obligatory.

As Lakoff describes the origin of the concept of freedom, his view is compatible with the idea that there can be societies which don't have a concept of freedom similar to the Western one inspired by liberal values. However, his view presupposes that every social group has some concept of freedom related to the concrete and empirical core concept developed on the basis of the individuals' experience of moving their own bodies. Having some concept of freedom, all social groups will therefore also make some kind of distinction between actions that are permissible (because they correspond to the ones a person should legitimately be free to perform), actions that are impermissible (because they invade the freedom of someone else) and actions that are obligatory (because they preserve somebody's freedom from being violated). The obligation of legal action as a means to protect the victim also beyond his/her will can be considered as a juridical translation of the psychological principle that we need to protect individual freedoms in any case. Which actions are considered permissible, impermissible and obligatory, will depend on the specific concept of freedom embraced by the judging subject. Which freedoms need to be protected most will also depend on the specific concept of freedom an individual has and therefore on the ICM that determines it. The more an action is perceived as absolutely permissible—i.e. is perceived as one of the fundamental freedoms people must enjoy—the more we will feel obliged to protect it and the more we will consider it impermissible to prevent someone from carrying it out.

Lakoff describes the mental structures that allow people to understand freedom as (IC)*Models* primarily because he thinks of them as a set of interconnected elements, while to build a model means to specify exactly the interconnections between a set of elements. However, these is also another reason why Lakoff's ICM should be called *model* and interpreted as a *model* that describes not only the concept of freedom of a specific person but also the way this person will produce moral judgments of permissibility, impermissibility and obligation. Indeed, even though both theories and models describe a set of structures and a set of processes that operate on these structures, models make these structures and processes more precise: they specify what the elements of the structures and of the processes concretely consist of.[9] *So, according to this definition, while Lakoff's general view on freedom is definitely a theory, ICMs are models of freedom*. The hypothesis which is implicit to Lakoff position and which this paper relies on is that in principle the structures and processes individual's ICMs are composed of *can indeed be precisely specified*.[10]

This conclusion has relevant consequences for the debate on moral judgment since, *if we could specify the ICM that determines an individual's concept of freedom, this would allow us to determine what action in what circumstances this individual will consider as permissible and why*. If impermissibility is described in terms of the actions that people are not free to perform because they violate the legitimate freedoms of others, then our concepts of freedom also decide which actions should be considered impermissible. Any disagreements can be traced back to the different ICMs that determine the individual's concept of freedom. The same applies to 'obligatory' which is interpreted in terms of the need to protect freedoms, i.e. to allow everyone to perform the actions he/she must be legitimately free to perform and to prevent everyone from performing those actions he/she should not be legitimately free to perform, because they violate the freedoms of others. What actions someone will consider as obligatory will therefore depend on the concept of freedom he/she has, i.e. on the model that specifies the content of the concept for him/her: describing that concept would allow us to determine the course of the moral judgment also as far as obligation is concerned.

[9] This definition of theories and of models resemble in a way that given by Thagard [38], even though for Thagard a model can only be computational, while Lakoff's model is not meant to be computationally realizable.

[10] Lakoff maintains that such a specification is possible in principle and gives important suggestions about how to develop it concretely. However, he doesn't offer any systematic representation of ICMs, so, the structure of ICMs hasn't been detailed yet.

4 Explanatory Advantages of Lakoff's View

Usually the idea of freedom people have and the moral judgments they produce are approached as two different issues. The view we put forward in the previous section suggests instead that they are connected and that we can use Lakoff's theory of freedom as a theoretical instrument to deal with the problem of moral judgment. In this part we analyze this connection from other points of view in order to show that Lakoff's theory has further relevant explanatory advantages. Specifically, it can help us explain some of the aspects of people's moral judgment that are usually considered problematic such as: (a) the fact that some moral violations are particularly serious—are, so to speak, more 'fundamental' than others; (b) the existence of infra-cultural similarity in moral judgment together with the variability of moral judgment within the same culture or group; (c) the variation of people's moral judgments depending on the situation and on the individuals involved; (d) and finally the fact that moral judgments are always related to emotions.

(a) In the preliminary part of this work we recall Turiel's (and colleagues') and Haidt's (and colleagues') view. According to Turiel the violations that people consider as strictly moral are *only* those relating to harm, rights and justice. On the contrary, Haidt suggests that—when we distance ourselves from the context of the western, educated and progressive society—we find that people consider not only these kinds of issues but *also* other issues, of a different kind relating to in-group/loyalty, authority/respect and purity/sanctity to be strictly moral. However, the fact that issues relating to harm, rights and justice are universally considered as strictly moral, while only some groups or cultures include in the moral domain also these other kinds of issues suggests that the first set of issues is the more basic one (about this see e.g. [9]. Moreover, the studies on the possible innate foundations of morality carried out e.g. by Pinker, Brown Mikaheil, Hauser etc. (see above §2) focus on a specific sub-group of the issues considered by Turiel: physical harm and murder as the extreme expression of harm. They implicitly assume that murder and harm are the most fundamental moral violations in the sense that they are cross-culturally perceived as the most serious ones. In fact, this position is compatible with the idea shared by a large part of the cognitive research on moral judgment that harm based violations have a special status and tend to be universally considered as morally wrong (see e.g. [3, 29, 30, 41]). These consideration allow us to sketch something like a scale of moral violations: issues related to harm, rights and justice are cross-culturally perceived as serious moral violations, however, among them murder and physical harm are perceived as the most serious violations; some cultures might also consider other kinds of issues as strictly moral, but the nature and the perceived seriousness of these violations is highly variable from one cultural context to another. We think that Lakoff's model of freedom can be of great help in explaining why harm and murder are perceived as the most serious moral violations and why some violation are cross-cultural while others depend on the cultural context.

Lakoff suggests that—in spite of its cultural and individual variability—everybody's concept of freedom shares a core concept: this core is the starting point for developing our abstract, adult concept of freedom and consists in the bodily experience of being free to move. Since the core concept is considered common for all human beings and therefore cross-cultural, it is plausible to assume that the violations of freedom which are more directly connected to the core concept will be cross-culturally considered as strictly moral violations. Moreover, if the core idea of freedom is first of all freedom to move, then "Harm (sufficient to interfere with normal functioning) is interference with freedom" in its more fundamental form [23]. Every hindrance to the normal functioning of the body—starting with physical elimination (as maximal interference with normal functioning), to physical harm, and to the prevention of the fulfillment of those primary needs (e.g. nutrition) which normal functioning depends on—will therefore be cross-culturally perceived as a very serious violation of freedom. Analogous considerations apply to coerced action [23] and to confinement, which are however considered less serious violations since they interfere with the freedom to act, but not with normal functioning, which is the very condition of the possibility of movement, at present and in the future.

In general, Lakoff's theory implies that the more a specific freedom is directly related to the core concept, the more its violation will be *cross-culturally* and *trans-individually* considered *highly* impermissible. Since physical elimination, physical harm and prevention from the fulfillment of primary needs for normal functioning rely on the core concept of freedom everyone shares, they will be unanimously perceived as the most serious violations of freedom and therefore as highly impermissible.

(b) Lakoff's theory of freedom can also help explain both the existence of infra-cultural similarity in moral judgment and the variability of moral judgment within the same culture or group. In general, members of a specific culture or group tend to produce uniform moral judgments because a culture or a group largely shares the same or analogous world views, while individual's subjective ICMs and frames are derived from these views, even though each person internalizes them in his/her own way. This subjective internalization of the cultural world views individuals are exposed to explains, on the other hand, why the moral judgments of members of the same culture or group can sometimes diverge. This way of explaining the cultural variability of moral judgment is particularly flexible because it does not only rely on the moral rules embraced by the group, but also on the specific concept of freedom a person has developed starting from his/her exposure to the group and from his/her personal reworking of the (politically and ideologically oriented) views on freedom he/she came into contact with. However, the flexibility allowed by Lakoff's theory is not absolute (free-floating), but assumes forms that can be determined (and described) in relation to the specific ICM of a person and of the frames that compose it.

This model is also compatible with the idea that a person's moral judgments might change over time and explains these changes on the basis of a structural revision of the previous ICM and frames of this person. Also in this case, the change is not free-floating, but depends on the specific elements of the ICM and frames that have been challenged and modified and on the specific connections between the elements of the ICM and frames that have been affected by this modification.

On the basis of Lakoff's model it is also possible to explain the differences in the application of moral rules presented by Dwyer using the notion of "schweeb". The world view of a culture determines which creatures are 'schweebs' (must be morally protected) and which are more/less 'schweeb' than others. Besides, world views also include (at least implicitly) some reasons to justify these choices. Since the individual ICMs and frames are internalized starting from the cultural world view the subject is exposed to, people will often inherit criteria used by their culture or group in determining who is a 'schweeb' and who is not (or who is more/less 'schweeb' than others). However, the distinction between 'schweeb' and 'non-schweeb' is not inherited as a dogmatic, singular rule, disconnected from any justification. Since this distinction is part of a structured world view, in which every assumption relies on specific arguments which support it, individuals can challenge it and—in case they are not convinced of its plausibility—they can work out justifications for alternative distinctions and therefore develop an individual ICM the principles of which diverge from those embraced by their culture. In turn, these justifications can sometimes also be accepted by others, become shared and modify a previous cultural world view. By being included in a complex ICM, the determination of which creatures are (more or less) 'schweebs' no longer appears to be the result of a meaningless, more or less contingent parameter that cannot be further investigated like the one Dwyer appeals to, but rather depends on the structure and the composition of individual ICMs and the frames they are composed of.

(c) Lakoff's view indirectly also offers some answers to the problem of explaining why the moral judgments of a subject can change depending on the concrete situation he/she is faced with and the people involved.

The view of freedom proposed by Lakoff has egocentric roots. Since our core concept originates from the natural and fulfilling experience of being able to move our own bodies as opposed to the experience of not being able to move as we wish or of being kept from moving, which is viscerally perceived as oppressive, overwhelming, frustrating and even painful, the concept of freedom we rely on is necessarily bound (also emotionally) to a first person perspective. Because of this, we can hypothesize that—when the permissibility, impermissibility or obligatory judgment does not concern abstract actors or situations (as often happens e.g. in an experimental setting), but rather concerns specific people whose actions have consequences for the judging subject—moral judgments are biased by our first person perspective on freedom and therefore by the tendency to protect our freedoms first. In fact, usually, in the case of conflict, the protection of our freedoms takes priority over the protection of others' freedoms and everyone is inclined to judge as impermissible those actions carried out by others that interfere with his/her freedoms, and to perceive as highly obligatory the protection of his/her freedoms (also to the detriment of the freedoms of others which fade into the background).

An extreme example to illustrate the point we want to make is the one, brought up previously, of illegal immigrants: when we judge e.g. whether it is permissible to turn back a boat of starving and exhausted immigrants who crossed a border illegally, the outcome of this judgment will depend on a two-step process. Firstly, our ICM and frames will lead us to conceptualize the immigrants in some way: we might come to conceptualize the immigrants primarily as human beings needing help, whose lives are in peril or alternatively as persons who can potentially threaten our freedoms. If we come to see them as people who can potentially threaten our freedoms (invade our territory, increase criminality, take our jobs, etc.), the egocentric factor will prevail and we will consider it permissible to turn them back, even though in the abstract we consider it impermissible to undertake actions that can potentially harm or kill other human beings. More generally, when an individual's ICM brings him/her to see another person as someone who can potentially threaten his/her freedoms, this will bias his/her judgment narrowing down the degree of moral protection assigned to this person. Actions that the judging subject considers absolutely impermissible when considered with respect to neutral or abstract situations, are judged permissible when they concern someone who might potentially threaten our freedoms.

Opposite considerations might apply when the judging subject has a positive relationship to one of the actors in the morally relevant situation. Indeed, there are cases where the opinion of the judging subject in a specific situation seems to diverge from his/her general moral principles and to be particularly indulgent in judging the actions of specific persons involved. At least some of these cases can be explained by appealing to the positive identification of the judging subject with specific actors. If he/she positively identifies or feels related to one of the persons involved in the situation, we might suppose that the egocentric bias of the moral judgment will influence his/her judgment in favor of this person: i.e. that the freedoms of the person he/she identifies with will take priority over the freedoms of other actors involved.

In general, whether we see a person as a terrorist or as a patriot, whether we consider someone a criminal or a contemporary Robin Hood depends on the specific ICMs and frames that determine our concept of freedom together with the concepts this is related to (such as 'justice', 'equality', 'human nature', 'competition', 'property', 'rights' and so on); however, once we categorize a person in a certain way—e.g. as someone potentially dangerous to us (as a criminal, as a terrorist, etc.); or as someone we identify with or who is positively related to us—then our judgment changes accordingly. While we are strongly inclined to protect the freedoms of the people we identify with or who are positively related to us (also to the detriment of the freedoms of other people), we tend to consider it permissible to limit in all possible ways other people's freedoms when we perceive they are in conflict with ours. The general idea behind this model is that we tend to be willing to protect the freedoms of others only as long as they don't come into conflict with our own. Therefore, Lakoff's view suggests that when we face some abstract moral dilemma with depersonalized actors, in our judgment we apply *in a neutral manner* the concept of freedom determined by our ICMs and frames. On the contrary, when we have better knowledge of the actors of a morally relevant situation, our judgments are strongly biased by how we think the actions of the people involved will affect us and by whom we identify with.

(d) Lakoff's model can also help explain another important factor in moral judgment which has remained implicit in the previous discussion: i.e. it allows us to account for the fact—widely discussed in the literature on morality—that moral judgments are always related to emotions and that people have strong feelings about their moral values (for reviews see e.g. [33], Chap. 1). Lakoff maintains that "freedom is felt viscerally, in our bodies, because, it is fundamentally understood in terms of our bodily experience": while the experience of being free to move is natural and fulfilling, the experience of not being able to move is oppressive, overwhelming, frustrating and painful. If this attempt to apply Lakoff's theory to moral judgment is correct, than the visceral emotional attachment everyone has towards his/her concept of freedom necessarily penetrates moral judgments since we will be viscerally attached to the idea of protecting our freedoms and preventing any actions that might prejudice them. In this way our emotional attachment to our moral values would be a direct consequence of our emotional attachment to freedom, independently of how we conceptualize it.

5 Concluding Remarks

The contemporary cognitive research on morality has mainly focused on cross-cultural and trans-individual moral principles and rules that can possibly be considered innate. This attention to the shared factors of moral judgment has led to disregard for the extreme variability of moral judgment which cannot be reduced to cultural differences only, but further concerns different individuals belonging to the same group or the same individual across time and in different situations. Some authors who—like Susan Dwyer—take this variability seriously, come to the conclusion that there are no commonalities among moral judgments as far as their outcomes are concerned and that everyone learns subjective moral rules. However, this view implies that it is impossible both to provide a model for moral judgment and to account for its cultural and individual variations, by explaining what they depend on.

In this paper we try to sketch a model of moral judgment that accounts for both the similarities and the differences. To achieve this aim the paper suggests first of all that we need to 'invert' the perspective commonly taken by the cognitive research on morality which focusses for the most part on those actions that are considered impermissible or obligatory, deriving the notion of permissibility from impermissible and obligatory: permissible actions are all those which are neither impermissible nor obligatory. The view we put forward here starts from permissibility and derives from it impermissibility and obligation. According to this model, people perceive as permissible those actions they think one must be legitimately free to do; they perceive as impermissible the actions that overstep the limits of one individual's freedom and encroach on the freedom of someone else and finally they perceive as obligatory the actions that protect personal freedom against invasion. The advantage of this inversion consists of establishing a connection between two different issues that are usually addressed separately: the debate on moral judgment on the one hand, and the problem of defining how we develop our idea of freedom on the other. More specifically, it allows us to link the issue of moral judgment with Lakoff's cognitive analysis of the concept of "freedom" (and of the understanding people have of freedom). The main aim of the paper is to show that Lakoff's theory indirectly offers a highly flexible and strong explanatory model of moral judgment. In fact, the paper points out that many aspects of moral judgment can be explained on the basis of the application of Lakoff's theory.

On the basis of Lakoff's theory we suggest that individuals' moral judgments depend on the concepts of freedom they have and these are, in turn, determined by their ICMs. Lakoff's perspective has many obvious limitations since it does not exactly specify the elements an individuals' ICM is made of and how/why ICMs vary from one person to another. In this sense Lakoff's view is first and foremost programmatic: it shows that our concepts of freedom must rely on structured models with strongly interconnected elements and it indicates a way to specify the precise structure of those models. Relying on a such a view, the hypothesis we put forward in this work also has a mainly programmatic aim. Indeed, it suggests that if we could describe an individual's ICM, we would be able to indicate which actions a person will judge as permissible, impermissible or obligatory and what this judgment is based on.

References

1. Arendt, H.: Eichman in Jerusalem: a Report on the Banality of the Evil. Addison-Wesley, London (1963–2006)
2. Bandura, A.: Mechanisms of Moral Disengagement. In: Reich, W. (ed.) Origins of Terrorism: Psychologies, Ideologies, Theologies, States of Mind. Cambridge University Press, Cambridge (1990)
3. Blair, R.J.R.: A Cognitive Developmental Approach to Morality: Investigating the Psychopath. Cognition **57**, 1–29 (1995)
4. Boudreau, T.E., Polkinghorn, B.D.: Reversing the Destructive Discourses of Dehumanization: A Model for Reframing Narratives in Protracted Social Conflict through Identity Affirmation. Res. Soc. Mov. Conicts Chang. **29**, 175–205 (2008)
5. Brown, D.: Human Universals. Temple University Press, Philadelphia (1991)
6. Chomsky, N.: Language and Problems of Knowledge: The Managua Lectures. MIT Press, Cambridge (1987)
7. Dellantonio, S.: Kognitive Module und moralische Kompetenz: Ist es möglich, der Moral eine biologische Grundlage zu geben? In: Fischer, M., Hengstschlger, M. (eds.) Genetic Screening. Ethik transdisziplinär Bd. 10. Peter Lang, Frankfurt a. M. (2010)
8. Dellantonio, S., Job, R.: Morality According to a Cognitive Interpretation. A Semantic Model for Moral Behaviour. In: Magnani, L., Carnielli, W., Pizzi, C. (eds.) Model-Based Reasoning in Science and Technology. Springer, Berlin (2010)

9. Dellantonio, S., Job, R.: Moral Intuitions vs. Moral Reasoning. A Philosophical Analysis of the Explanatory Models Intuitionism Relies on. In: Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science. Springer, Berlin (2012)
10. Dellantonio, S., Pastore, L.: Teorie morali e contenuto cognitivo. cognitivismo, postmoderno e relativismo cultural. In: Meattini, V., Pastore, L. (eds.) Individuo, identità, soggetto tra moderno e postmoderno. Mimesis, Milano (2009)
11. Dellantonio, S., Pastore, L.: 'Libertà' senza significato? Concetti astratti, cognizione e determinismo linguistico. Rivista Internazionale di Filosofia e Psicologia **2**, 164–186 (2011)
12. Dwyer, S.: How Good is the Linguistic Analogy? In: Carruthers, P., Laurence, S., Stich, S. (eds.) The Innate Mind. Culture and Cognition, vol. 2. Oxford University Press, Oxford (2006)
13. Haidt, J.: The Emotional Dog and its Rational Tail: a Social Intuitionist Approach to Moral Judgment. Psychol. Rev. **108**, 814–834 (2001)
14. Haidt, J., Bjorklund, F.: Social Intuitionists Answer Six Questions about Moral Psychology. In: Sinnott Armstrong, W. (ed.) Moral Psychology. Vol.2. The Cognitive Science of Morality: Intuition and Diversity. MIT Press, Cambridge (2008)
15. Haidt, J., Graham, J.: When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. Soc. Justice Res. **20**, 98–116 (2007)
16. Haidt, J., Graham, J.: Planet of the Durkheimians, where Community, Authority, and Sacredness are Foundations of Morality. In: Jost, J., A. C. Kay, H.T. (eds.) Social and Psychological Bases of Ideology and System Justification. Oxford University Press, Oxford (2009)
17. Haidt, J., Joseph, C.: The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Cultural-specic Virtues, and Perhaps Even Modules. In: Carruthers, P., Laurence, S., Stich, S. (eds.) The Innate Mind, vol. 3. Foundations and the Future. Oxford University Press, Oxford (2007)
18. Harman, G.: Explaining Value. Oxford University Press, Oxford (1999)
19. Hauser, M.D.: Moral Minds. How Nature Designed Our Moral Sense of Right and Wrong. Collins Publisher, New York (2006)
20. Hooks, G., Mosher, C.: Outrages Against Personal Dignity: Rationalizing Abuse and Torture in the War on Terror. Soc. Forces **83**, 1627–1646 (2005)
21. Kelly, D., Stich, S., Haley, K.J., Eng, S.J., Fessler, D.M.T.: Harm, Affect and the Moral/ Conventional Distinction. Mind Lang. **22**, 117–131 (2007)
22. Lakoff, G.: Women, Fire, and Dangerous Things. What Categories Reveal about the Mind. University of Chicago Press, Chicago (1990)
23. Lakoff, G.: Whose Freedom? The Battle Over Americas Most Important Idea. Picador, USA (2006)
24. Lakoff, G.: The Political Mind. Why You Cant Understand a 21st-Century Politics with an 18th-Century Brain. Viking, USA (2008)
25. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York (1999)
26. Merskin, D.: The Construction of Arabs as Enemies: Post-September 11 Discourse of George W. Bush. Mass Commun. Soc. **7**, 157–175 (2004)
27. Mikheil, J.: Rawl's Linguistic Analogy. Ph.D. Thesis Cornell University Press, New York (2000)
28. Mikheil, J.M.: Law, Science and Morality: A Review of Richard Posner's 'The Problematics of Moral and Legal Theory'. Stanf. Law Rev. **54**, 1057–1127 (2002)
29. Nichols, S.: Mindreading and the Cognitive Architecture Underlying Altruistic Motivation. Mind Lang. **16**, 425–455 (2001)
30. Nichols, S.: Sentimental Rules. On The Natural Foundations of Moral Judgment. Oxford University Press, Oxford (2004)
31. Nucci, L.: Education in the Moral Domain. Cambridge University Press, Cambridge (2001)
32. Pinker, S.: The Language Instinct. Morrow, New York (1994)
33. Prinz, J.J.: The Emotional Construction of Morals. Oxford University Press, Oxford (2007)

34. Prinz, J.J. Is Morality Innate? In: Sinnott-Armstrong, W. (ed.) Moral Psychology. Evolution of Morals, vol. 1. MIT Press, Cambridge (2008)
35. Smetana, J.G.: Understanding of Social Rules. In: Bennett, M. (ed.) The Development of Social Cognition: The Child as Psychologist. Guilfort Press, New York (1993)
36. Smetana, J.G.: Context, Conflict and Constraint in Adolescent-Parent Authority Relationships. In: Killen, M., Hart, D. (eds.) Morality in Everyday Life: Developmental Perspectives. Cambrigde University Press, Cambridge (1995)
37. Smetana, J.G., Rote, W.M., Jambon, M., Tasopoulos-Chan, M., Villalobos, M., Comer, J.: Developmental Changes and Individual Differences in Young Children's moral Judgments. Child Dev. **84**, 683–696 (2012)
38. Thagard, P.: Mind. Introduction to Cognitive Science. MIT Press, Cambridge (2005)
39. Tisak, M.: Domains of Social Reasoning and Beyond. In: Vasta, R. (ed.) Annals of Child Development, vol. 11. Jessica Kingsley, London (1995)
40. Turiel, E.: The Development of Social Knowledge. Cambridge University Press, Cambridge (1983)
41. Turiel, E.: The Culture of Morality. Social Development, Context and Conict. Cambridge University Press, Cambridge (2002)
42. Turiel, E., Killen, M., Helwig, C.: Morality: Its Structure, Functions, and Vagaries. In: Kagan, J., Lamb, S. (eds.) The Emergence of Morality in Young Children. University of Chicago Press, Chicago (1987)

Models of Moral Cognition

Jeffrey White

Abstract This paper is about modeling morality, with a proposal as to the best way to do it. There is the small problem, however, in continuing disagreements over what morality actually is, and so what is worth modeling. This paper resolves this problem around an understanding of the purpose of a moral model, and from this purpose approaches the best way to model morality.

1 Introduction

> *The process here analyzed is not a dream, a fancy floating in the air; it is perfectly real, and by no means infrequent.*
>
> —Schopenhauer[1]

A model is a representation of salient aspects of a system that, when rendered together, articulate an essential function in a more efficient way than the original, a replica or a duplicate. So, models are created for reasons other than for the creation of one of these other things. Some models are explanations. For example, a model of disease represents how pathology progresses. Some models are made to help realize an original. For example, models of buildings inform architects and engineers how to make original buildings which, once constructed, can serve in the creation of duplicates or replicas. Models of this sort are especially important when new answers are necessary, novel creations in response to new problems and the questions that these raise. This paper is interested in models that do this sort of

[1] [1], p. 170.

J. White (✉)
KAIST, Daejeon, South Korea
e-mail: jeffreywhitephd@gmail.com

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_20, © Springer-Verlag Berlin Heidelberg 2014

work, but rather than help in building better houses, the models that we are after should help us to become better people. Rather than model new places to stay, new futures to grow into, ourselves included.

Two general forms of moral model are prevalent, and both seem to aid moral development. The traditional form is one of narrative and ethical theory expressing principles affirmed by intuition and enculturation through example, demonstration, and argument, and the other, more recently popular form is that of mechanistic and information processing models of specific subroutines and circuits within the brain, within the organism, or within the extant ecosystem, all working together to tell the story of morality. Which mode of representation is best?

Twenty years ago, anticipating the impact of the cognitive sciences on moral philosophy, Stephen Stich asked a similar question, and pointed in the direction of psychological representations. A quick review of his reasons for this will help to provide some context for the rest of this paper, as well as set up some important issues to be met with along the way, including the role of models in moral practice, and potential for future research.

2 Looking for Mr. Goodmodule?

In a talk from 1989 published in 1993, Stephen Stich argued that a central project in traditional moral philosophy had been chasing its tail, and issued a sort of rallying call to future-minded moral philosophers around a forecast that “the beginnings of moral philosophy fall squarely within the domain of cognitive science”[2] [2]. Stich argued that moral philosophy had been searching after things that “do not exist,” and he identified a set of “Platonic assumptions” responsible for leading the inquiry astray. The first problem was that some philosophers had “presumed that the mental structures underlying moral judgments are rather like definitions” in that they “specify individually necessary and jointly sufficient conditions for the application of moral concepts.” The second problem was the claim to reliable intuitions about these definitions, with the “central strategy in testing a proposed definition” being merely “to compare what the definition says to what we would say about a variety of actual and hypothetical cases.” And, the third assumption that Stich found active was mistaking the central task of moral philosophy to be “making explicit the necessary and sufficient conditions that, presumably, we already tacitly know,” (pp. 3–4) Thus, we see moral philosophy setting out for itself both the terms of its own inquiry and the standards for their evaluation. Self, chasing, tail.[3]

[2] [2], p. 14. Noted pagination belongs to the author’s copy, a copy of which is maintained by Joshua Knobe online at the address cited.

[3] It is as I had read the other day, “It is a familiar problem in recent philosophy that to the extent my experience of another person can be assimilated to ready-made experiential categories, I have

Rather, Stich saw the future of moral theory in psychological alternatives “that do not involve necessary and sufficient conditions.” These aim to represent moral concepts in forms that people already comfortably employ in directing and evaluating everyday morally *in*significant action, like “the knowledge structures that guide our expectations in reading stories about restaurants and other common social situations.”[4]

Stich found that these everyday frames, as well as other systems of representation under psychological consideration—“Mathematical knowledge, knowledge of various sciences, and common sense knowledge in various domains” (p. 13)—are analogous to moral systems in a very important way, in that people

> *can offer a complex, subtle, and apparently systematic array of judgments about particular cases, with little or no conscious access to the mechanisms or principles underlying these judgments* (pp. 13–14).

This fact sheds some light on the purpose of moral philosophy, as well as on the structure of moral judgment. Moral judgment is the product of something deeper, that informs consciousness. And, the best ways to represent morality are those ways that best communicate the significance of these deeper things. Stich approached these issues through his primary vocation, as an ethics teacher. Given the purpose to effectively communicate moral concepts, truths, so that students can assess, assimilate, and critically evaluate morally salient situations, thereby empowered through understanding to a lifetime of free philosophical self-determination, the best way to represent morality is easily determined. In the same ways that people demonstrate, learn and understand morality, already, through direct and indirect experience of the moral lives of self and others:

> *Exemplar models of conceptual representation, and more sophisticated variations on the theme that invoke “scripts” or stories, also suggest an explanation for the fact that those engaged in moral pedagogy generally prefer examples to explicit principles or definitions. Myths, parables, fables, snippets of biography (real or fanciful)—these seem to be the principal tools of a successful moral teacher. Perhaps this is because moral knowledge is stored in the form of examples and stories. It may well be that moral doctrines cast in the form of necessary and sufficient conditions are didactically ineffective because they are presented in a form that the mind cannot readily use* (p. 11).

An exemplar is a “specific instance” of some unique thing “falling under a concept.” On the view that concepts are represented in the form of exemplars, “categorization” of perceived objects

> *proceeds by activating the mental representations of one or more exemplars for the concept at hand, and then assessing the similarity between the exemplars and the item to be categorized* (p. 10).

(Footnote 3 continued)
not really gotten beyond myself. Rather, in the experience of the apparent other, I have merely reconfirmed or reconnected with a prior sense of self-identity.” [3], p. 119.

[4] Today, this is deeply researched, with some technology employed in reading minds by reproducing field effects within areas of the brain matching those of the donor.

In this way, exemplars serve as vehicles for moral knowledge by demonstrating modes of being through which moral concepts are expressed. But more than that, exemplars are a special kind of model, for they are something that one can "model" himself after. One can direct one's life along similar paths as those demonstrated by exemplars, mimicking routine actions in routine contexts, and one can compare one's self against exemplified demonstrations as standards. In the comparison, one feels what it is like to differ from these examples, feeling the difference as the satisfaction or failure to meet exemplified standards. How well exemplars work applied to novel situations, however, is another problem, and one that we will come to as this paper closes.

The space of academic philosophy offers the leisurely reflection necessary for ready analysis of possible situations and application of principle, where exemplars are not such efficient vehicles of knowledge.[5] This is clearly not the most efficient way to model morality, however, unless expecting everyone with moral aspirations to spend their days engaged in professional moral philosophy. Rather, exemplar models work in everyday life because moral knowledge is about human lives, and human life is more effectively represented in examples and demonstrations than categories and principles.

Since the time of this writing, great progress has been made toward confirming Stich's forecast that "the beginnings of moral philosophy fall squarely within the domain of cognitive science." By 2000, Nancy Eisenberg was able to report that "Philosophers' changing view of the role of emotion in morality is consistent with the predominant view of emotion in psychology today" in understanding that "higher-order emotions such as guilt and sympathy are believed to motivate moral behavior and to play a role in its development and in moral character" ([4], p. 666). And, since 2000, the area between moral philosophy and the cognitive sciences has exploded, with disciplines at its core notably absent from Stich's short list of philosophy, anthropology, and cognitive psychology. To these must be added a cluster of new fields falling directly under his forward gaze, experimental philosophy, neurolaw, neuroethics, neurophenomenology, and social cognitive neuroscience amongst them, all with a focus on correlating "what it feels like" with neural activity understood either metabolically or computationally. All of this confirming Stich's:

> *strong suspicion that progress in understanding how people represent and use moral systems will not be made until scientists and scholars from these various disciplines begin to address the problem collaboratively. Indeed, one of my goals in writing this chapter is to convince at least some of my readers that it is time to launch such a collaborative effort* ([2], p. 14).

Here is a short list of traditional philosophical terms that are being naturalized through ongoing interdisciplinary work around the issue of moral cognition:

[5] And, as aging studies have shown, older people tend to rest on old ideas, with aging lazy philosophers, hashing out the fine points of established definitions is expected according to brain research.

- "experience" as "conditions under which associations are formed between novel stimuli and biologically innately significant events, typically innate triggers," ([5], p. 656)
- "intuition" as product of one thread of the dual-processing portrait, "associative" and "attuned to encoding and processing statistical regularities, frequencies, and correlations in the environment," ([6], p. 990)
- "moral intuition" as "fast, automatic, and (usually) affect-laden processes in which an evaluative feeling of good-bad or like-dislike (about the actions or character of a person) appears in consciousness without any awareness of having gone through steps of search, weighing evidence, or inferring a conclusion," ([7], p. 998)
  - with the "key functional difference" between moral and other intuitions being "that moral intuitions appear to make a difference, directly, to how we act and react," ([8], p. 7)
- "moral emotion" as an extension of root-level survival circuits distributed throughout the body and realized in the brain as emotions that are at once evaluating and motivational, [5]
  - with Jonathan Haidt confining the moral to just those emotions that are concerned with others rather than with one's own prudential self [9].
- and most recently "conscience," "a neural process that generates emotional intuitions combining somatic perception (the gut reaction) with cognitive appraisal concerning a special subset of goals"([10], p. 156).

When Stich was writing, without models of neural processing assembled from basic neurological research, it had been easier to conceive of universally binding rational principles than similarly effective sets of somato-affective markers and their corresponding motivations. Traditionally, intuitionist, sentimentalist, or emotivist accounts of moral cognition had been hamstrung by a limiting capacity to draw their subjects in clear and distinct terms. Now, the "new synthesis" in neuroethics promises to open new avenues to toleration, compassion, and mutual understanding built on what is best understood as the "shared body."

Not confined to individual human agency, neurological research has also informed thinking on the issue of collective agents, where mirror systems and empathy embodied in individual subjects help to explain inter-subjective associations whereby "Some people may act "as-if" a certain belief was their own without actually endorsing it themselves," with the result being the appearance of unity, and so of collective agents as entities in their own right.[6] Thusly, through advances in functional imaging, a real-time picture of man's moral reality built of affect, bottom-up, is being extended from neural substrate to intuition to institution

[6] [11], p. 336. Such tendency to social coherence is also affirmed in the cognitive "switch" that turns individual fans into a seething mass, helping to explain the loss of self also experienced by persons caught up in the mass psychology of crowds.

and social organization, deep in territory traditionally belonging to moral philosophy.

In this spirit, Young and Saxe point out that individual differences in moral judgment can be mapped onto regularly recurring patterns and intensities of activity in different areas of the respective subjects' brains. These differences correlate with education, upbringing, and routine attitude, and even characteristic mood, with Saxe reminding us, for example, that "people who are generally disgusted make harsher moral judgments of unrelated incidents." Their approach is to discover such patterns of activation common between individuals and groups, thereby revealing the "independent psychological components of moral judgments" and the neurological basis for "apparently arbitrary "cultural clusters" of moral value." Ultimately, Saxe, suggests that mapping neural differences between parties to moral differences, "may help us to understand and resolve moral disagreements not only between individuals but also on a broader scale." She, as Stich two decades prior, points to the future of moral inquiry in psychological representations, forecasting that "The next stage for research must therefore be to understand the structures underlying these differences" ([12], p. 324).

Pursuit of the mechanisms underlying moral judgments may reveal a universal basis for moral judgment in these same mechanisms, with the hope that these provide all that is necessary for moral guidance. Consider Jonathan Haidt's assessment of the relative importance of intuition and moral reason in that effort:

> *In other words, evolution shaped human brains to have structures that enable us to experience moral emotions, these emotional reactions provide the basis for intuitions about right and wrong, and we* (*or, at least, many moral theorists*) *make up grand theories afterward to justify our intuitions* ([9], p. 68).

And, Cokely and Feltz second this sentiment, suggesting that not only are these theories post hoc, but they may also be counterproductive:

> *In an uncertain and complex world such as ours, we should not expect or necessarily even want to always be governed by processes that maintain logically coherent cognition* ([13], p. 358).

This is a long way from Stephen Stich rejecting necessary and sufficient conditions as necessary and sufficient for moral theory. In the words of Darcia Naevaez, "the pendulum is swinging in the other direction and reasoning is often considered unnecessary" ([14], p. 164).

It may be that remaining rational is not always rational. And, understanding the grounds for moral differences through somato-affective mechanisms is a long way from the high point of the rationalist pendulum in the other direction. But does distance equate with progress? Rather, it is my strong suspicion that progress on the issue of moral representations cannot be made unless the highpoints of either are reconciled with one another, seconded by an even stronger suspicion that we have been in this situation, before.

As an example of a rationalist high point in moral theory, consider the following conclusion from Hastings Rashdall on the plausibility of intuitionist theories that morality is an emotion:

> *I have tried to suggest to you that they can be met in as purely a scientific and dispassionate manner as that in which they are (at least sometimes) defended. But the scientific spirit does not require us to blind ourselves to the practical consequences which hang upon the solution to not a few scientific problems. And assuredly there is no scientific problem upon which so much depends as upon the answer we give to the question whether the distinction which we are accustomed to draw between right and wrong belongs to the region of objective truth like the laws of mathematics and of physical science, or whether it is based upon an actual emotional constitution of individual human beings, which may once have possessed, and possibly may still possess, a certain survival-value in the evolution of the species to which those individual belong. That emotionalist theory of ethics however little intended to have that result by its supporters, is fatal to the deepest spiritual convictions and to the highest spiritual aspirations of the human race* ([15], p. 199–200).

For Rashdall, the problem with intuitionism is not what it tells us about human beings as a product of evolutionary forces beyond their control. Rather, morality is about that part of human evolution that people do control. It sets out ideals, "the highest spiritual aspirations of the human race" in certain terms. What is valuable now is determined on the basis of these ideals, rather than how evolution has shaped us to feel about it. Constructs of human reason, theories and hypotheses, abductions, principles, expressions of "objective truth" like those of mathematical and physical law, tell us what is valuable beyond the range of our evolved capacity to feel about things.

This line of thought represents a strong counter to the nativist push to write reason mostly out of the moral chain of causation. Intuitionists, on Rashdall's account, fail to adequately weigh the consequences of the action. They account for the motivation, the antecedent. But, without objective means to weigh ends of action against one another, when morally salient emotions conflict, it is impossible to decide on the relevant course of action, "for it is impossible to pronounce one motive higher than another in the abstract, without reference to circumstances" ([16], Chap. 4). And, there is no guarantee, or at least not guarantee enough, that evolution has prepared us for the circumstances that confront us at any given moment.[7]

[7] Consider this adaptation of a famously mistaken line, from another famous and oft mistaken utilitarian, John Stuart Mill—The only way that we can know what is worth seeking or avoiding is because these are actively sought or avoided. On Mill's account, you would be better off Socrates suffering then follow the nose of evolution to the end of history. Because, without some power to determine the situation into which life places a human being, that human being remains a slave, thus failing to qualify for moral consideration, at all. Emotions remain stuck in the situation as it is, and insofar as humanity binds its sights to an emotional moral mooring, regardless if these have an evolutionary basis, it is as if mankind had never crawled from the primordial muck, leaving behind its correspondent morality, and adapted to the world as it should be. On the other hand, moral reason attaches to morally ideal situations and principles, because it aims at the best possible consequence regardless of how we feel about it. It is not what evolution has brought us to, but what we do with who we are, today, and tomorrow, these are the ethically

In this light, consider Peter Singer's position, that the contribution to moral philosophy from the cognitive sciences may be negative, confirming only those aspects of morality that should be pared away in pursuit of adequate moral theories. On Singer's assay, only moral skepticism is the alternative to "the ambitious task of separating those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational basis," with the full intention of discarding all those without a rational basis ([17], p. 351). And, so far as neuroethicists over-confidently swinging the theoretical pendulum are concerned:

> *Advances in our understanding of ethics do not themselves directly imply any normative conclusions, but they undermine some conceptions of doing ethics which themselves have normative conclusions. Those conceptions of ethics tend to be too respectful of our intuitions. Our better understanding of ethics gives us grounds for being less respectful of them* (p. 349).

It is not what evolution has brought us to, but what we do with who we are, today, and tomorrow, these are the ethically relevant aspects of moral life worth talking about. Any evolved, innate emotional dimensionality may describe what we do on the basis of how it feels, but it does not tell us what should be done, regardless, and it is unlikely that revealing the structures further underlying moral reasoning is going to do so, either. Intuitions, insights there into and their theoretical offspring, are merely imperfect starting points to responsible moral agency, and those who hold innate processes as upper and lower limit to the space of moral theory are at best misinformed and at worst naive.

As expressions of our highest, most distinctly human capacities to conceive of ourselves, our world, and the world that we leave behind after actions right or wrong accumulated, these rationalist constructions pull us forward, rather than push us along. They tell us why we live, not just why it feels good or bad when we do it this way or another. And this is their purpose. They open up the space of goal-oriented categories, allowing a currently bad feeling to be endured for a better one. Without these goals, and especially without their development into philosophical ideals, there is no possibility for the analysis of consequences.

(Footnote 7 continued)
relevant aspects of moral life worth talking about. What decides between the emotions is the purpose, the rationally constructed ideal end and object of the action and the emotion that wins is the one that brings about the best possible moral situation consonant with that action's purpose. Any evolved, innate emotional dimensionality may describe what we do on the basis of how it feels, but it does not tell us what should be done, regardless, and it is unlikely that revealing the structures further underlying moral reasoning is going to do so, either.

3 Two Moral Templates

All of the evidence points to the fact that “Morality is a natural phenomenon. No myths are required to explain its existence” ([17], p. 337). And this clarity extends to all levels of human conduct, with Jonathan Haidt asserting that “Moral systems are interlocking sets of values, practices, institutions, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make social life possible” ([9], p. 70).

The issue is what we do with this understanding, not only to make social life “possible” but to make it better. One way in which this already happens involves tempering immediate desire for long-term cooperative goals. Likewise, Darcia Narvaez warns against the reduction of moral motivation to intuition and emotion due to the limits of “gut-reaction” assessments in both picking out and assigning adequate significance to morally salient features of complex and changing situations. Narvaez points out that morality requires an individual “to step away from his own interests and from current norms to consider more inclusive and logically just forms of cooperation” ([14], p. 167) utilizing all forms of information available in the construction of moral ideals and principles that help us to work together toward more just arrangements.

The ability to create and to set out for one’s self moral ideals and ideal situations, better situations, as well as to empathize with others, taking up their situations as if one’s own, “in their shoes” so to speak, is moral imagination. Lorenzo Magnani and Emmanuel Bardone characterize moral imagination as “analogical and metaphorical reasoning” that is “very important” to the practice of ethics “because of its capacity to “re-conceptualize” the particular situation at hand,” representing the situation as it should be or could have been [18]. Building from work done by Magnani (2001), they suggest that analogical reasoning is a type of model-based reasoning. That being so, moral imagination sets out situations to be sought and others to be avoided, based on information from one’s own and others felt, expressed, embodied situations [19].

Building from work done by Magnani, (2007), Magnani and Bardone note another way in which model based reasoning sheds light on moral cognition [20]. Ends set out and achieved may be worked toward without something like what Stich noted earlier in terms of other forms of knowledge, without “conscious access,” with agents remaining able to execute sophisticated patterns of behavior, along the “how/that” distinction in epistemology generally. Magnani and Bardone review the notion of “tacit templates” to account for “embodied, implicit patterns of behavior” ([18], p. 100) that are essentially context specific routine actions either non-reflexively triggered through prior training to “be selected from those already stored in the mind–body system, as when a young boy notices his baby sister crying and, without thinking, automatically tries to comfort the infant,” or “*created* in order to achieve certain moral outcomes” (authors’ emphasis, p. 100). This process of developing a model routine and internalizing it in self-direction, toward some further goal, is an illustration of the constructive role of what Magnani has developed as "moral mediators."

Specifically, the sort of model that we are after here is an example of a "task-transforming" external representation. This kind of representation simplifies an otherwise complex task by transforming "difficult tasks into ones that can be done by pattern matching," thereby making possible solutions to problems at hand "transparent," with the understanding that "The more transparent the agent makes the task, the easier it is to find the proper solution" (p. 103).

In this section, we will look to two candidate sources for the sort of task-transforming representations necessary.

First, let's look at Jonathan Haidt's "social intuitionist model." Haidt defines moral intuition as a capacity to realize moral truth without an exercise of reason, but rather through motivating emotions, with the content of these intuitions including emotional valences on the model of perception, with the shape of these valences ultimately due to evolution, recognizing that "it is very difficult to create a fear of flowers, or even of such dangerous things as knives and fire, because evolution did not 'prepare' our minds to learn such associations" ([21], p. 58). Supporting these evolved moral processes are moral modules, "small sets" of which are productive of moral intuitions, and Haidt and Joseph posit the existence of four fundamental sets of modules concerned with purity, reciprocity, hierarchy and suffering. Paying special attention to that concerned with purity, Haidt and Joseph paint a compelling portrait of the extension of moral principle from innate neural structure, providing a universal basis for morality grounding the common forms and functions of moral principles active in different cultures, regardless of apparent differences:

> *Over time, this purity module and its affective output have been elaborated by many cultures into sets of rules, sometimes quite elaborate, regulating a great many bodily functions and practices, including diet and hygiene. Once norms were in place for such practices, violations of those norms produced negative affective flashes, that is, moral intuitions* ([21], p. 60).

The social intuitionist model has "four links" ([22], p. 818). These proceed stepwise as follows. First, the "intuitive judgment link" by way of which "moral judgments appear in consciousness automatically and effortlessly." Second, the "post hoc reasoning link" "in which a person searches for argument that will support an already-made judgment." Third, the "reasoned persuasion link" in which a person communicates his moral reasons to others, and may persuade others by "triggering new affectively valenced intuitions in the listener." Finally, the "social persuasion link" is a passive mechanism potentiated by human sensitivity to "group norms" such that "the mere fact that friends, allies, and acquaintances have made a moral judgment exerts a direct influence on others, even if no reasoned persuasion is used" most notably to agree with allies and friends and to regard others vice versa, resulting in social cohesion through a mechanism not unlike that detailed in Magnani, 2011 [23].

The "central claim" of Haidt's nativist model is that "moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, post facto moral reasoning" ([22], p. 817). Moral reasoning is defined narrowly as "conscious mental activity that consists of transforming given information about people in order to reach a moral judgment." The social intuitionist model "gives moral reasoning a causal role in moral judgment but only when reasoning runs through other people" because "reasoning is rarely used to question one's own attitudes or beliefs" ([22], p. 819). Haidt defends this hypothesis partly on the basis that challenging comfortable prior evaluations, judgment, or beliefs is resisted due to the fact that these re-evaluations threaten existing self-conceptions and world-views according to which life is interpreted as meaningful and on the right track. This leads reason to the exercise of self-defense, as if a "lawyer" rather than a "scientist," either with different object notions of truth. He also cites memory bias, status-quo and self-interest to motivate a "make-sense epistemology" in which "the goal of thinking is not to reach the most accurate conclusion but to find the first conclusion that hangs together well and that fits with one's important prior beliefs" ([22], p. 819).

Consistent with the "social persuasion link," in which an exemplar, prototype, or demonstration of a moral judgment may lead others to follow suit, Haidt and Joseph assert the superiority of the virtue theoretic approach over other approaches to moral development in that "it sees morality as embodied in the very structure of the self, not merely as one of the activities of the self," with virtues themselves represented as "social skills" "closely connected to the intuitive system," the possession of which are evidenced by "the proper automatic reactions to ethically relevant events and states of affairs"[8] ([21], p. 61). Moreover, as the criteria according to which moral action and moral character are commonly evaluated are virtues relative culture and practice, Haidt and Joseph suggest that we take advantage of the body's "preparedness" to make some associations that expedite learning about those things over others.

But, how to take advantage of this preparedness? Haidt and Joseph differentiate their approach from traditional virtue ethics according to a relative de-emphasis of cultural-environmental determinations of virtue, and increased emphasis on "a smaller number of phenomena that are located more in the organism than in the environment," at once recognizing the central importance of each moral agent's unique embodied situation in the instruction of moral virtue through the inculcation of appropriate "flashes" of moral intuition:

> *These flashes are building blocks that make it easy for children to develop certain virtues and virtue concepts. For example, when we try to teach our children compassion, we commonly use stories about mean people who lack those virtues. While hearing such stories children feel sympathy for the victim and condemnation for the perpetrator. Adults cannot create these flashes out of thin air; they can only put children into situations in which these flashes are likely to happen* ([21], p. 63).

[8] The inverse of which being Thagard's "situational distortions."

Ultimately, the placement into morally instructive situations, so that innately present moral processes are attuned to salient moral dimensions otherwise lacking in experience, is the limiting factor in the growth and development of moral agency. Of course, Haidt's portrait recalls Rousseau's famous farmer's plot demonstration in *Emil*, and as well suffers from a singular objection. It is not so easy constructing these situations.

Furthermore, picture the eventuality of generation after generation putting selves and children into situations that feel, from a common evolutionary basis, like the right situations to be in. How is this different than the arbitrary hand of nature circling the thread of human fate back on itself? From whence does the hero arise who breaks this cycle and frees the future from the past?

It is the task of the exemplar to demonstrate this sort of information that is impossible to represent otherwise. These sources of moral knowledge do produce those flashes of understanding, while also contributing information about timing, and fine motor action, as well as affective cues signaling appropriate motivations and social cues. However, Socrates is more than two centuries dead. Christ, Mohammed, the great leaders in Martin Luther King, Jr., and Gandhi, all dead and lest we wait for exposure to a possible hero, it is up to us to stand in for absent exemplar. Consider, again that history is a circlet, nose to tail. And, we are back at the beginning, rehashing the same old controversies.

Should we wish to encourage the origination of such moral exemplars, potentiated by moral models, to change the world, is this the most effective way to do it? Where the task is moral self-regulation and philosophical self-determination, a successful moral mediator in the form of a tacit template for moral becoming must simplify this process while at once making solutions to everyday moral problems transparent to the subject. Though such illustrations as Haidt's do render solutions transparent, they do nothing to make them easier to reproduce. This means that they are hard to use, represented in ways that people cannot readily employ. If, indeed, facilitating moral agency is the goal of a moral model, then it is difficult to see how nativist mechanisms can further this goal.

After all, it is impossible to expose a child to the all the necessary right things at the right times, and for many, moral life is mostly a series of corrections on what had been a childhood full of bad information. If we take this notion seriously, and we should, then the direction that Haidt's model is taking us begins shed light on the possible source of a standard for moral worth around the exercise of available agency to re-direct and refine the given moral life.[9] This is an expression of virtue, if it is possible at all.

[9] On Martin Heidegger's account, a person does not choose where and what and who he or she is. Rather, he is "thrown" into a situation, and is left to come to terms with it, to discover it and to understand it, and to courageously become what is necessary to take up his thrown condition, its history, and its people, and employ what potential he can to moving that situation forward, toward a morally ideal situation consonant with an essentially social yet individually embodied condition, all while confronted by its inevitable conclusion in death. A person is governed by moods, with mastery over moods necessary for moral freedom, especially mastery over the dread

If we allow that it is possible, then this, the "honest toil" of moral agency rests not in availing to hardwired precepts, but rather in moral education, self-development and self-regulation, with the leverage points to affect this process most often outside-in and top-down. That is, it typically requires leisure, self-reflection, and a good bit of luck to borrow from Aristotle. Accordingly, one might object to this briefest of presentations of Haidt's intuitionism on the grounds that self-reflection, the thinking part, is not given due consideration, after all.

Further evidence against intuitionist accounts of morality might also be derived from research proposing that a specific morally motivating emotion does not exist. Batson, for example, locates what others consider moral motivation in selfish gain through "moral hypocrisy," the successful presentation "as-if" being moral without motivation to become so [24]. However, this position stands against some prima facie evidence to the contrary. If moral motivation is limited to selfishness, how does a moral ideal present itself, at all, let alone universally? Is it that becoming a moral exemplar is simply an ultimate realization of hypocrisy, pursued for its presumed social and material benefit, universally realized and sought after? Given the tragic ends having met many memorable moral exemplars across the cultural-historical continuum, and the inspiring pro-social influence their examples continue to have on people around the world, this seems unlikely.

A better answer to these questions may be found in the universal structure and function of moral cognition involving the integration of intuitive and rational mechanisms into the unified prospective concerns of a morally self-regulating entity and fundamental unit of moral value, a structure understood traditionally as conscience.

Space forbids adequate review of the philosophical tradition around conscience. It had been a cornerstone in ethical theory until the late 20th century. Conscience has all but disappeared from moral theory, except for medical ethics where the freedoms of doctors and health care professionals to deliver or to restrict medical attention, care, while constrained by law and business policies that may run contrary to those freedoms, remains a contested issue. In this field, Donald Sulmasy offers a "contemporary" view of conscience deserving brief review here.

Echoing Rashdall's assessment of the importance of our understanding of morality, Sumasy holds that it is "impossible to suggest anything more important to the moral life than conscience." On Sulmasy's account, both individuals and institutions are beholden to conscience, with conscience representing

> *the most fundamental of all moral duties—the duty to unite one's powers of reason, emotion, and will into an integrated moral whole based upon ones most fundamental moral principles and identity* ([25], p. 138).

(Footnote 9 continued)
angst of death. Depending on how far from an ideal situation one is "thrown," more or less work must be done to correct for poor moral upbringing during adulthood in striving toward that morally ideal situation on which his identity rests. See, Heidegger takes Aristotle's "a friend is another me" and makes it a fundamental part of the human condition, *mitdasein*. The other is not another me. The other is me.

According to Sulmasy, conscience has two aspects, one “turned toward its origin” and the other “turned toward moral acts.” It comes to our attention when “deliberating about particular cases.” It “establishes a felt need” to act according to “fundamental moral commitment to act with understanding” in a way that maintains moral integrity, by resulting in a situation consistent with personal moral precepts. The established feeling constitutes an evaluative “meta-judgment” over the situations brought about through action, both prospective and retrospective, in the form of guilt or shame associated with unsatisfactory ends, and with peaceful wholeness and integrity the reward for having done the right thing, and having brought about the right end.

Approaching the topic of conscience from the philosophy of psychology and cognitive sciences, Thagard and Finn refer to conscience as “the internal sense of moral goodness or badness of one’s own actual or imagined conduct,” as a “kind of moral intuition, and as “an indicator of the legitimacy of a moral judgment,” bridging innately grounded affect and “internal and external standards” while informing us “about what our moral goals are, as well as about good ways to meet these goals” ([10], pp 150, 168, 161, 161, and 163). This description explicitly unifies “top” and “bottom” processes, with conscience working bottom up in producing what Haidt’s model accommodated as emotional valences, this one on the order of rightness and wrongness.

Thagard’s model rests in an understanding that “emotions are both cognitive appraisals and somatic perceptions, performed simultaneously by interacting brain areas” (p. 151). Cognitive appraisals are judgments on “the extent to which something aids or hinders our goals.” Somatic perceptions are “perceptions of bodily states.” Their combination results in a view of cognition that evaluates possible goals in terms of anticipated body states.[10] Conscience, expressed as guilt and shame, thus expresses a situation arrived at in violation of some other emotional valence,[11] and these are not limited to social feelings. Rather, Thagard recognizes the fact that moral and non-moral situations elicit activity in similar regions of the brain, suggesting that there is “nothing special about the brain processes involved in moral intuition compared to emotional consciousness in general.” Conscience however, on Thagard and Finn’s estimation, concerns moral goals only, such as “increase the well-being of people in general,” “act in accord with abstract moral principles such as fairness and respect for autonomy,” and “satisfy the expectations of social groups such as family and comply with religious standards or other moral code” (p. 153). This is of course to beg the question—Is it conscience that delineates the moral from non-moral?—but we shall leave this question behind.

In short, conscience, when judging an action right, is expressed as a positive emotional valence associated with the satisfaction of the goal toward which that

[10] Note the parallel with Sulmasy’s two dimensional characterization.

[11] We may also deduce that the voice of conscience is anticipated guilt or shame for some situation made possible by some entertained action.

action aims. Working against these goals results in negative emotions. Thus, conscience represents a mechanism for social compliance, as well as motivations to some other goal for which some positive valence is associated.

When it comes to moral self-regulation and instruction, rather than to conscience, directly, Thagard points to his "informed intuition" model for moral problem solving.[12] This four-step model is decidedly top-down, proceeding thusly:

1. Set up the decision problem carefully. This requires identifying the goals to be accomplished by your decision and specifying the broad range of possible actions that might accomplish those goals.
2. Reflect on the importance of the different goals. Such reflection will be more emotional and intuitive than just putting a numerical weight on them, but should help you to be more aware of what you care about in the current decision situation. Identify goals whose importance may be exaggerated because of emotional distortions.
3. Examine beliefs about the extent to which various actions would facilitate the different goals. Are these beliefs based on good evidence? If not, revise them.
4. Make your intuitive judgment about the best action to perform, monitoring your emotional reaction to different options. Run your decision past other people to see if it seems reasonable to them (p. 162).

This model stands in contradiction to Haidt's hypothesis that "reasoning is rarely used to question one's own attitudes or beliefs." This is a decision procedure seeking reflective equilibrium through a critical evaluation of how given beliefs contribute to the realization of deliberate goals, calling for their revision on this practical basis. Contrary to the intuitionist program, Thagard's takes care to set out ideal situations and evaluate the feelings that arise in their respective consideration, and this gives a critical stance from which to weigh the rationality of given emotional valences. Thus, Thagard's decision procedure goes a long way to answering objections to intuitionism leveled from the likes of Rashdall while remaining sensitive to motivating emotions, and opening the decision process to others who may be affected by actions in question.

But, is this the best way to represent morality in order to further the purpose of moral models, facilitating moral becoming? It certainly stands as an improvement, of sorts, over the virtue approach in that it can be applied in the consideration of hypothetical situations under one's own self direction.

Thagard's decision procedure breaks free from affective limits, and right that it should. Due attention must be given to what constitutes morality in addition to affect, specifically sources of moral freedom rather than evolved pre-determination.[13] The effective and affective detachment from immediate environmental

[12] Which may in moral cases perhaps be called the "educating your conscience" model

[13] After all, if we are not free to determine for ourselves what is right and wrong, and further to act toward one and away from the other, then any talk of morality rapidly reduces to pharmacology.

pressures is a source of human freedom, with this capacity archetypically realized as syntactical, symbolic, “offline” processing consistent with the perceptual basis of symbols and linguistic representations. (c.f. [26]) In this dimension, Thagard’s approach to informing moral intuition is on point. However, it is difficult to identify advantages of Thagard’s over other heuristics in framing moral problems, such as decision trees and reflective equilibrium approaches.

It is tedious, requires special time and attention to execute outside of the flow of everyday life, and even if beneficial given leisure, it fails to give direction in how to frame moral problems in a way to best inform moral intuitions. Rather, likely due to the view that there is no special set of morally specific modules in the brain, and no specifically moral processes in cognition, moral problems are approached as any other. In every case, required processing is slow, and so not suitable to directly inform some situations, but rather is best employed in reflection during moments of relative leisure, to recall an opportunity dear to Socrates, in order to rehearse and potentiate “the proper automatic reactions to ethically relevant events and states of affairs.”[14] As a result, it ultimately fails to render the process of becoming a moral person transparent.

However, it is clear how Thagard’s serves as a compliment to Haidt’s approach. Ideally, then, a model intended for moral self-development and instruction would marry the approaches of Thagard and Haidt, while taking advantage of embodied moral processing in a way that facilitates moral becoming through making the process of self-transformation, itself, transparent.

4 The Worm and the Mollusc

> *Although science likes to separate component processes for closer analysis, sometimes this gives the wrong impression—as if one can truly separate the person from the situation, reason from emotion, or intuition from unconscious reason.*
>
> —Darcia Narvaez[15]

First, I think that we can begin to make sense of continuing disagreement over the source and shape of morality through two observations and an image.

In my experience, people view cognition in ways that reflect their own cognitive styles, and cognitive styles are forged by the specific character of, and tempered by the breadth and depth of experience. Philosophers spend a lot of time thinking, while others may spend relatively more of their time doing. As philosophers are most often employed as educators, thus, we find people who spend their

[14] In other words, moral autonomy is to be found in the application of reflective analysis and moral imagination towards the preparation of innate capacities to feel, judge, and act, i.e. in the practice of traditional, especially Socratic, ethics.

[15] [27], p. 185.

time thinking concerned with communicating moral truths to others who more likely spend their time doing. The more that we reflect on emotions, after all, the sooner they are categorized, and it is easy enough to see how, at least in a man's understanding, this pattern of action might coopt an otherwise elephantine emotional life. But, there is no use in it if the elephant isn't frenzied and restless. There is no sense asking "What it feels like" of an analytic moral philosopher, if the philosopher has never felt it. And, if he has never felt it, then what he has left are his categories and conditions, which is where, I think, we started off with Stich in this paper's introduction.

I further suspect that some disagreement over the nature of morality is due to the subtle abuse of the common conception that human neural processing is dual in nature. Involving

> *two distinct systems through which human beings apprehend reality: System 1 is emotional, affective, intuitive, spontaneous and evolutionary prior; System 2 is rational, analytical, reflective and occurred later in our evolution* ([28], p. 175).

Along with this distinction has arisen a torrent of inquiry into the neural substrates of moral processing that has grown increasingly philosophically sophisticated, and controversy has arisen as these inquiries are framed and results interpreted providing physiological bases for phenomena which had been, traditionally, the domain of moral philosophers. As the theoretical reach of neurology into traditional moral philosophy has deepened, conflict has arisen between theorists who take morality and moral agency as an essentially rational exercise in self-determination, a "System 2" or "top-level" product, and those who take it as a product of evolved processing extended from basic operations maintaining physical integrity in the face of changing environmental pressures, as an essentially affective, "System 1," rather than rational activity. Champions of these respective approaches have contrasted their positions in very strong terms, and this has resulted in controversy. Controversy, moreover, that is not new, and perhaps requires not repeating.

Finally, consider that people have tended to regard the ways in which humans differ from other animals rather than their similarities as the locus of moral value, just as they have for reason and consciousness, categorically defining other animals exempt from moral consideration. It is my suspicion that this sort of reasoning, so "intuitive," has contributed to a misunderstanding of morality that remains implicit in attributions of moral value today. There is more in common between caterpillars and human beings than between human beings and most of the rest of the materials in the universe. Is it possible that some of this common structure is crucial to the moral structure of human beings, as well?

Consider the following story from the life of naturalist Jean Henri Fabre as related by Robert Kirkman, about a type of social caterpillar called a "pine processionary." These caterpillars "venture from the shelter of the nest" at night, in single file lined up without gaps, with "each caterpillar adding a strand of silk to the trail laid down by the leader." One day, Fabre looped this thread back on itself. And, as Kirkman quotes,

> *The unbroken chain eliminates the leader with his change of direction; and all follow mechanically, as faithful to their circle as are the hands of a watch. The headless file has no liberty left, no will; it has become mere clockwork* ([29], p. 27).

They followed in circles for a week. Such life, for a human being, may not seem worth living. There must be more than that, and it is to the difference between human and caterpillar that people have tended to look, with the implication that morality is not on the model of a caterpillar.

But, why?

Who can say that they have not been in the position of those caterpillars, perhaps once, following friends, associates, lovers on courses that only left them spent, lost, hungry and a week behind?

Consider Haidt's portrayal of the embodied condition as a small stick rider atop a massive elephant, ostensibly the driver but vastly overpowered and at the whim of the beast [30]. This illustration represents a correction on the presumption that people are essentially rational agents, and it does something more. It advises how to most effectively direct one's emotionally grounded life. The trick in directing one's life is to get the rider and the elephant working together.

This model has obvious advantages over, say, dualist models. For instance, there is prima facie reason to take good care of the emotional vehicle that is the body, where for the dualist the body may be more limitation than empowering transport. And, it does capture a sense of what it feels like to be a human being in a humorous way that is easy to accept and employ. However, it does not seem to reflect Sulmasy's profile of a Janus-faced mechanism for moral meta-judgment. And, as for our goal of best representing morality for moral development, what Darcia Narvaez calls "moral self-becoming," it is difficult to see how Haidt's illustration can be of much use.[16]

Let's start out for a better representation by returning to the beginning of the paper, to make something of Stich's use of Barsalou in positing more than one mode of representation at work in moral cognition, with

> *the mental representation of "goal derived" categories, such as things not to eat on a diet… may have a format that is quite different from the mental representation of apple, fruit, or dog.*

The implication is that a good model of morality may need to represent morality in more than one way, corresponding with different mental capacities and modes of operation. This characterization feeds Sulmasy's description of conscience, as well, with moral goals associated with or derived from principled moral conviction and the qualification of other representations falling under these goals colored accordingly. Further, according to Sulmasy, "conscientious" persons may change goals upon "learning certain empirical facts," ([25], p. 144) thereby

[16] And, besides, there is a troubling aspect to Haidt's image. This is that there is a man in the position of reason, and this reveals a tacit association between Haidt's conceptions of humanity and of reason that might be taken to locate moral value in reason.

educating conscience through a directed search for and exposure to such facts as seen in Thagard's informed intuition model.

Haidt's stick-figure elephant rider also represents two modes of representation at work. But, this picture does nothing to clarify the processes that tie these modes together, not in a pro-moral, or in any other way unless one wishes to carry the metaphor of stick rider further—"Be good to your elephant, and your elephant will be good to you," and so on.

Consider, rather, moral cognition on the model of an emotional inchworm ridden by an information processing bivalve. One, the inchworm, reaches one end of itself forward to feel out possible new situations, while the other end remains rooted in the original. Once felt, the bivalve can open to this information in order to determine what being in that situation would confront it with. On this image, there is no separate human rider struggling atop some furious beast. Rather, we have a single organism of two processes, one reaching forward or back to possible situations, and the other processing available information to compare with the still retained original. The inchworm feels out new situations, while the mollusk comes to terms with them. And the end selected is the one that feels best in the terms reached.

This model of cognition represents the dual nature of cognition in a way that these processes are active in the discovery of the world, in the generation of new experience, and also opens avenues to discussion of virtues, such as courage, versus vices such as recklessness, in a very clear manner. In terms so simple as to invite skepticism, courage requires that one come to terms with the situation that he seeks through action. Without this process, the agent is reckless, and ultimately immoral.

Before detailing this image further, let's examine the model of moral cognition from which it arises, the ACTWith model.

The ACTWith model was originally conceived of as a model of philosophical conscience, informed by Ron Sun's CLARION model of human learning [31]. "ACTWith" stands for "As-if Coming-to-Terms-With," representing a processing framework composed of a four-fold cycle that may be pictured as a sort of intuition-reason pump on the model of the human heart, with the heart being traditionally the embodied locus of conscience. The cycle proceeds as follows:

1) As-if (open) coming-to-terms-with (closed)
2) As-if (open) coming-to-terms-with (open)
3) As-if (closed) coming-to-terms-with (open)
4) As-if (closed) coming-to-terms-with (closed).

Open processes gate information into a process, closed operations process that information, with the open "as- if" operations feeling a situation out, and the open "coming-to-terms-with" operations defining the situation accordingly. So, for instance, in the closed/closed mode, the agent may act on the basis of interred information, returning to the open/closed mode, whereby the newly acquired situation after action is first felt out, and so on through the cycle. Similar processing occurs in active compassion and empathy. Feeling out another's situation is facilitated by affective and effective cues which provide comportment information and permit their direct embodiment through mirroring of that embodied condition.

In ACTWith notation, during the o/c stage of processing the agent opens to the situation. In the o/o, the agent feels as if in that situation while opening existing terms of understanding to revision on the basis of this new information, and during the third stage, c/o, the agent updates his existing understanding, not only feeling as if in another situation but understanding this as fact. In this mode, an emotionally laden conception of a candidate situation is generated, and this portrait is compared with the original, with the felt difference between them constituting motivation to seek or to avoid that candidate. During the c/c phase, the agent may act toward that situation, or return to the process of farming for more and better ones.

Different cognitive styles arise through the routine commitment of cognitive resources to the different modes of information processing, with the habitual embodiment of these modes in certain types of situations resulting in the development of different personalities and prejudices. Allocation of resources may be conceived of in terms of clock cycles, electrochemical potentials, or simply as time spent engaged in a certain mode of processing. For example, as the relative evaluation of other situations equally means one's own or another's, an agent may be habitually open to his own possible situations (o/o) while remaining indifferent to those of others (c/c). Regarding feeling out another's situation (o/c–o/o), if one's moral cognitive routine commits ample resources to identifying, recognizing, and personally realizing signs of affective and effective states, then this contributes to a certain cognitive style, including the projection of moral archetypes and the emergence of moral exemplars. In particular, the habitual exercise of the o/c/- o/o modes in morally significant situations potentiates exemplary kindness as well as wisdom, due to the fact that experiential resources are rapidly expanded, and bases for common understanding and terms for communication expanded, all contributing to a decidedly pro-social personality type. This cognitive style is "conscientiousness."

The ACTWith model makes easy sense of other basic moral attitudes, too. In compliment to Stich's "Platonic assumptions," consider the following "Socratic precepts" that arise from normal ACTWith operation.

The first of these assumptions is "Know nothing." Socrates was famous for suggesting that, though confirmed the 'wisest man in Athens,' he knew nothing. His method in discovery through discourse involved always beginning with the situation as understood by his interlocutors, and proceeding from there towards an adequate assay of the matter at hand. On the ACTWith model, this is represented by the first steps. In meeting with others, Socrates opens to the situation, then opens to the terms to which they have come in determining the situation, only feeling out and assessing further possible situations after this preliminary stage. By this precept, thus, one must adopt a situation as if one's own in order to begin to know why it is or is not satisfactory, why movement from this position (literal and figurative) is necessary, in order to lead from there to something better. Prior experience is active beginning in the third stage if this Socratic method is modeled after, but starting open to "what it feels like" to be in other situations, and informing one's understanding on this basis without prejudice is key. Making this movement habitual is the first step in becoming a conscientious moral agent.

A second Socratic precept is "Never cross your daemon." Socrates was famous for refusing to aid in the arrest and eventual execution of Leon of Salamis, and also for saying that he was gifted with an innate sense of justice, a "daemon" that forbid him from doing the wrong things. All that he had to do, he told us, was not to cross his daemon in order to emerge the 'most just man in Athens.' This function of conscience is represented in the ACTWith model as follows. As the cycle of processing completes, with terms of understanding come to insofar as resources had been dedicated to their assay during the first stages, the c/c stage draws the agent in on itself in preparation for action. Here, the infamous "voice of conscience" may arise, barring action and so barring passage to associated situations. Here, the last and the future situations are held together, at once, by either end of the illustrative inchworm, at the moment that the inchworm may commit, lifting its tail from its prior situation to pull itself forward into the next. Anticipating that chosen end, updating information until the commitment to the new situation is enacted, conscience reveals that progress to this new situation will result in a loss of progress toward some internalized moral ideal self-representation. That is, one feels as if he will no longer be his own best example of life worth living because the agency that results in said situation is contrary to the sense of agency exemplified in one's "highest spiritual aspirations," to become the best person one can possibly become. In this final instant before action, with both situations bridged and the embodiment of the new situation imminent, the agent is confronted by what Kant would call "self-repugnance" or self-disgust at the self that results from this situation. Thus, it is not the end, or the action itself, that are rejected in the "veto" of conscience, but rather what is rejected is the self that one will become through said action and at said end. This characterization captures the way in which conscience associates with integrity, feeling of "wholeness" and self-esteem, in natural, easy to employ terms.

The preceding Socratic precepts represent a traditional understanding of conscience while presenting this understanding in a way that is both consistent with what is understood about the neurology of moral cognition and that takes advantage of what is known about these processes in order to facilitate the self-direction of these processes towards a unifying purpose, moral self-development. These emerge from normal exercise of the ACTWith model. The ACTWith model, moreover, is able to accommodate different accounts of moral cognition, as well, even those that seem contrary to the model itself. These other accounts of moral cognition can be informatively mapped onto the ACTwith operations, showing that the ACTWith model is more fundamental.

Consider the following passage from Adam Smith's *Theory of Moral Sentiments* as he describes the process whereby he comes to understand the moral significance of another's embodied condition. Standard ACTWith notation has been added:

By the imagination we place ourselves in his situation [*O/C*], *we conceive ourselves enduring all the same torments* [*O/O*], *we enter as it were into his body*

[*O/O*]*, and become in some measure the same person with him* [*O/O*]*, and thence form some idea of his sensations* [*C/O*]*, and even feel something which, though weaker in degree, is not altogether unlike them* [*C/O*]*. His agonies, when they are thus brought home to ourselves* [*C/O*]*, when we have thus adopted and made them our own* [*C/C*]*, begin at last to affect us, and we then tremble and shudder at the thought of what he feels* [*O/C - > C/C, in reflection*] ([32], Sect. 1.1.2).

Similarly, Thagard's guide for informed intuition can also be mapped onto the ACTWith model. And, though Thagard's is not primarily a model of moral cognition, in so far as it is applicable to moral direction it should proceed according to the ACTWith logic if the ACTWith model is successful in articulating a universal structure for moral information processing according to which other approaches can be relatively evaluated and recommended. ACTWith notation and brief interpretive comments are added, as follows:

1. Set up the decision problem carefully. [O/C]—feel out the space of possibility.
2. Reflect on the importance of the different goals. [O/O]—attune one's self to the likely realization of different possibilities.
3. Examine beliefs about the extent to which various actions would facilitate the different goals. [C/O]—refine preconceptions based on expected outcomes.
4. Make your intuitive judgment about the best action to perform, monitoring your emotional reaction to different options. [C/C]—act towards a new situation, then/or repeat the cycle.

And, we can do the same thing with Haidt's four-step social intuitionist model, too:

1. The "intuitive judgment link" by way of which "moral judgments appear in consciousness automatically and effortlessly" is O/C wherein arise gut-reactions to possible situations.
2. The "post hoc reasoning link" "in which a person searches for argument that will support an already-made judgment" is C/O, as terms of understanding are farmed for confirmation of the gut-reaction product of step 1. Note that Haidt effectively skips the O/O step, wherein new terms of understanding are generated, bottom-up, so the C/O stage is rather anemic on Haidt's model, thereby limiting moral development consistent with his presumption that reasons is not part of the chain of moral causation.
3. The "reasoned persuasion link" in which a person communicates his moral reasons to others, and may persuade others by "triggering new affectively valenced intuitions in the listener" is C/C, as the persons perform communicative acts, effectively changing the social dimensions of the situation. Presumably, then, the person will enter into a new cycle of processing from this altered situation, until action toward the realization of the felt goal is potentiated.
4. Finally, the "social persuasion link" representing the "direct influence on others" that morally salient action exerts, "even if no reasoned persuasion is used" seems to be a complex of O/C (open to the demonstrated examples of others), O/O (being directly influenced to follow or to reject those examples),

C/O (exemplars represent a mode of understanding, with this understanding applied to like situations), and C/C (actively exemplifying virtue or vice as information for others).

It is not troubling that these processes are not replicas or duplicates of the ACTWith model, as they each express different assays of moral cognition consistent with the cognitive styles of their creators. It is merely a sign that the ACTWith model is more fundamentally sound than these others in that the ACTWith model had been designed in order to be able to accommodate these variants, as well as more radical variants such as those demonstrated by psychopaths, both individual and institutional, as well as artificial moral agents, and examples from traditional moral philosophy [33–37].

Some comparisons are in order. There is nothing essentially moral about Thagard's model for informed intuition. Neither is there anything essentially moral about Haidt's "nativist" model. One is an extension of individual prudence endorsed through friendly confirmation in the final step. The other is an extension of primal mechanisms aiming at contextually various satisficing conditions, with moral excellence arising through some unspecified mechanism (though perhaps in the ACTWith spirit due to the projected emotional fit of the organism to some projected ideal moral situation). On the other hand, the ACTWith model is essentially a model of morality. On its account, cognition essentially sets out and weighs potentially embodied situations, not simply one's own and not neglecting that potentials can approach zero. This is all undertaken in energetic terms which, due to common physiology and natural law, provide a universal basis for the relative evaluation of embodied situations, and so provide a universal basis for the moral judgment over any given situation and the actions, conventions, and institutions that bring it about.

Space forbids further details, but, very quickly, perhaps the greatest upshots to this model are the following.

One, it encourages the development of moral exemplars, helping to draw human moral development forward. And, it does this while making consistent sense of ongoing research in moral cognition. For example, the ACTWith model makes sense of recent research that persons who are generally or easily disgusted exhibit harsher moral judgment than others less sensitive to disgust, and that these results can be reproduced when the evaluative basis in mood is temporarily induced through disgusting and irritating noises.

Two, the ACTWith model naturalizes intention in an intuitive and useful way. With conscience understood as the felt comparison of relatively well-ordered situations with the ideally ordered case understood as an ideal arrangement of objects on minimal dimensions, a "-science," and with the felt tension between situations motivational, intention can be understood as "in-tension." Given the common energetic basis of the ongoing analysis of situations on the ACTWith model, intension is understood as the internal, motivating and relatively evaluative felt strain, or "tension," between conscientiously compared situations, reference to which expresses both the motivation to some end as well as the end, itself. This interpretation falls in well with everyday language. For example, one "intends" to bring a situation about simply because it is a better situation to be in according to the terms of evaluation brought to bear in the comparison, noting that these terms need not be subject to conscious selection.

One may object that this makes no sense of intentions over individual objects. I think that such a possible objection is mistaken for two reasons. One, there is no compelling evidence that cognition attends to individual objects rather than possible effects that these objects may have on possible situations. One need only consider how dramatically a situation can change when it includes a door key, or a restroom, to see that, as individual objects in the placements and properties change, so do the situations in which they take part. And, moreover confirmed in intuition, the only sense in which these objects do take place, or not, is that in which the situation as a whole is transformed by their presence or lack thereof.

Another upshot for the ACTWith model is that the ACTWith program naturalizes freewill as the embodied metabolic potential to posit, alter, construct and to otherwise act toward ends of one's own self-determination, not least through attending to and altering the weights attached to salient terms brought to bear in rational analysis. Most importantly, this process underwrites philosophical self-determination, the particular capacity of directed thought to affect the sort of person that one will become through action by inculcating automatic or practiced reactions to specific opportunities when so presented. Ultimately, this capacity is due to the fact that thinking about one situation rather than another, in one set of terms rather than another, expends similar amounts of physiological potential, leveling the decision space given relative lack of urgency. Though fundamental to Thagard's informed intuition model, this aspect of moral agency is discounted on Haidt's, but only in the ACTWith model is the metabolic basis for cognition as well as bodily actions rendered in one coherent frame.

Finally, the ACTWith model helps to make sense of otherwise troubling concepts from the philosophical tradition concerning moral self-development, encouraging the aspiration to moral ideals rather than wrote internalization of moral principle or affect, and this deserves the briefest of accounts. By the ACTWith program, conscience signifies the enveloping framework of cognition, guiding an agent from situation to situation. It lays out possible ends of action as situations in which the agent innately seeks to retain integrity by maintaining equilibrium between internal and external forces, and this embodied logic, along with embodied limitations, allows for their comparison and relative evaluation, with differences providing motivation to move toward some and away from others. Fundamental terms for the relative evaluation of situations are derived from metabolic, physiological constraints, and are thus essentially energetic rather than material. Conscience so conceived is the felt comparison of situations in the constant adjustment of any dynamic agent to its changing internal and external environments, in the human instance via homeostatic regulation of embodied processes extending through moral cognition, including the comparison of possible situations hypothesized in terms with which the person already cognizes and acts as made available through limiting experience, i.e. "moral imagination."

As the constitution of these hypotheticals proceeds from a limited sphere of individual experience, augmented by affective and effective mirroring as well as taught "top-down," there is great potential for the scope of conscience to expand over the course of operation. As terms increase, given sufficient resources, the agent may develop capacities to simultaneously evaluate greater numbers of dimensions and to more readily identify morally salient dimensions. With the space of action mapped through this operation properly understood as metaphysical, rather than merely physical, conscience motivates the agent to seek situations with minimal strain between one's own and others' current and expected future situations, with the global minimum—informed as described, through habitual conscientiousness—specified as the Kantian "summum bonum" [35].

This inspirational quality is obvious from the ACTWith structure. According to Kant, an agent would be merely "a marionette or automaton" without the tension between the sensible and the ideal, with any sense of freedom a "mere delusion," freedom "only in a comparative sense, since, although the proximate determining causes are internal, yet the last and highest is found in a foreign land" ([38], p. 102). Substitute "pine processionary" for "marionette" and the relationship becomes clearer. After all, should a marionette live, it is not a life worth living, perhaps even less so than the caterpillar's and for similar reasons. The source of the motivating moral tension ultimately drawing the moral agent on to the Kantian "kingdom of ends," aspiring to Kantian reverence and away from moral repugnance, is conscience as understood on the ACTWith model.

5 Conclusion

I want to close by reconsidering Rashdall's phrase, introduced earlier, that "the scientific spirit does not require us to blind ourselves to the practical consequences which hang upon the solution to not a few scientific problems."

How we conceive of morality has practical consequences. These conceptions leave morality more or less available to practice. So, conceptions that make solutions to moral problems transparent are the best.

Perhaps the most important moral problem confronting every moral agent is who he will become through a life of action, a good person or bad. The ACTWith model helps to make solutions to this ongoing problem transparent. Moreover, potentiating moral self-determination raises the bar of human leadership, and this is promising for the future of human tolerance and liberty, qualities sadly failing to tyranny in the current era. After all, who willingly serves a lesser man than himself, to lesser ends than he is able, but a slave, or a worm, or a marionette, all without moral significance? This answer to this question is also rendered transparent on the ACTWith model.

"Ultimately, a genuine leader is not a succor for consensus but a mold of consensus" [39]. Leaders do more than make and break laws. They exemplify ways of life, ways which, due to the nature and namesake of their positions, others follow, a fact of the human condition to which Haidt gives due attention. Towards these, and for example in "conscientious objection" radically different ends, the ACTWith model facilitates life-long moral development in a practical, holistic way, being an intuitive, quick and transparent heuristic, which, easily employed routinely and habitually entrains the agent into a specifically moral virtue, conscientiousness. In short, where other models ask if an act is prudent, or safe, if it feels good or even if it is popular, the ACTWith model of conscience asks of the proposed end of action, "Is it right?"

For John Dewey, the capacity to imagine other situations, to manipulate those situations, and to relatively weigh them, as is required in assessing the consequences of actions, "constitutes an extension of the environment to which we respond" ([40], p. 387) Imagination confronts the thinker with possible situations, by placing the thinker in those situations, forcing the thinker to come to terms with those situations as if they were his own.[17] This is because cognition is not separate from the body and from its situation. Rather, in Dewey's words, "mind is a complex function of the doings and under goings of encultured, embodied, historically situated organisms, continuous with physical systems" ([41], p. 10).

This understanding, nearly a century old, is worthy of claiming today. And this reveals something about the tradition in moral philosophy and the future of moral theory. Though the cognitive sciences have contributed to our deepening understanding of the wheels that turn within us, it has offered less in the way of self-regulatory powers over those same processes. Intuitions and their evolutionary origins do not directly show us how to succeed in becoming moral, to remain so, or to aspire to some higher level of moral virtue. Moreover, such a neurologically based understanding of morality is not easily applied in the evaluation and similar reform of institutions and collectives, themselves by some regarded as morally significant individuals in their own right. As well, neurological models are useful, but by no means prescriptive in considerations of the engineering design and moral standing of artificial moral agents, or any other morally significant entity, individual or collective, so far beyond study. The ACTWith model of moral cognition was developed to overcome these shortcomings.

In the end, our deepening understanding of embodied moral mechanisms may not be the most important tool in our moral development. And this returns us to the inspiration that set us out on this journey, Stich's call to collaboration on the most important questions in moral life. With Stich, in the beginning of this paper, we found moral philosophy chasing its own tail, without the influence and information from other disciplines, especially psychology. Here, at the end of our discussion, do we not find the cognitive sciences chasing its own tail? After all, in testing for morally salient functionality specific to certain areas of anatomy, do the scientists not test from the same set of action potentials and expectations that guide their

[17] This portrait is supported by evidence that similar pathways of neural processing "are activated both during prospection and during hypothetical moral decision-making," ([40], p. 749) and that all cognition is essentially of the embodied condition.

own subjective experience? They confirm, then, only themselves in what they study. Their work reflects their evaluations, and expectations, as these are all that they know to challenge. But, what of moral ideals? Where are these to be tested, weighed, measured? Is it not from philosophy, and not cognitive science, that any question as to the potential realization of this human body arises? And without this view to the human future, what is the value of anything, at all, but what it is rather than what it might become?

With these questions in mind, let's close with some reflections on the future on moral philosophy from Young and Koenigs. Though they show no doubt that extra-rational processes play decisive roles in moral judgment, for better or for worse, given that "A coarse summation of the clinical findings is that individuals who exhibit abnormal emotional processing also exhibit systematically abnormal moral judgment," these scientists note that, perhaps, the pendulum of progress into the question of moral representations has reached its zenith in the cognitive sciences. They tell us that "Even though the acquisition or expression of moral knowledge may be a suitable subject of scientific inquiry, science cannot reveal what is morally right or morally wrong," and that the "brain may thus constrain the moral mind, but how we decide to deal with such constraints may be best determined in philosophical debate." Finally, looking forward, they point back to moral philosophy, and back in the direction from which we have come. Their advice is to "return to the likes of Kant, Hume and Mill or join the efforts of a new camp of scholars, empirical philosophers, who seek to marry descriptive and normative approaches to human moral psychology" ([42], p. 77). Advice worth following.

References

1. Schopenhauer, A.: The Basis of Morality. (Translated with introduction and notes by A.B. Bullock.) Swan Sonnenschein & Co., London (1903)
2. Stich, S.: Moral philosophy and moral representation. In: Hechter, M., Nadel, L., Michod, R. (eds.) The Origin of Values. Aldine de Gruyer, New York (1993). http://www.unc.edu/~knobe/x-phi/stich.pdf
3. Andersen, N.: Conscience, recognition, and the irreducibility of difference in Hegel's conception of spirit. Ideal. Stud. **35**(2), 119–136 (2005)
4. Eisenberg, N.: Emotion, regulation, and moral development. Annu. Rev. Psychol. **51**, 665–697 (2000). http://psych.colorado.edu/~tito/sp03/7536/eisenberg_2000.pdf
5. LeDoux, J.: Rethinking the emotional brain. Neuron **73**, 653–676 (2012)
6. Osman, M.: An evaluation of dual-process theories of reasoning. Psychon. Bull. Rev. **11**, 988–1010 (2004)
7. Haidt, J.: The new synthesis in moral psychology. Science **316**, 998–1002 (2007)
8. Kauppinen, A.: Intuition and belief in moral motivation. In: Björnsson, G. (ed.) Moral Motivation: Evidence and Relevance. Oxford Univ. Press, Oxford (in press). http://tcd.academia.edu/AnttiKauppinen/Papers
9. Haidt, J.: Morality. Perspect. Psychol. Sci. **3**, 65–72 (2008)
10. Thagard, P., Finn, T.: Conscience: what is moral intuition? In: Bagnoli, C. (ed.) Morality and the Emotions, pp.150–169. Oxford University Press, Oxford (2011)

11. Krause, J.: Collective intentionality and the (re)production of social norms: the scope for a critical social science. Philos. Soc. Sci. **42**, 323–355 (2012)
12. Young, L., Saxe, R.: Moral universals and individual differences. Emot. Rev. **3**(3), 323–324 (2011)
13. Cokely, E.T., Feltz, A.: Adaptive variation in judgment and philosophical intuition. Conscious. Cogn. **18**, 356–358 (2009)
14. Narvaez, D.: Moral complexity: the fatal attraction of truthiness and the importance of mature moral functioning. Perspect. Psychol. Sci. **5**, 163–181 (2010)
15. Rashdall, H.: Is Conscience an Emotion? Three Lectures on Recent Ethical Theories. Houghton Mifflin, Boston (1914)
16. Rashdall, H.: The Theory of Good and Evil: A Treatise on Moral Philosophy. Oxford University Press, London (1924)
17. Singer, P.: Ethics and Intuitions. J. Ethics **9**, 331–352 (2005)
18. Magnani, L., Bardone, E.: Distributed morality: externalizing ethical knowledge in technological artifacts. Found. Sci. **13**(1), 99–108 (2008)
19. Magnani, L.: Abduction, Reason, and Science. Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
20. Magnani, L.: Semiotic brains and artificial minds. How brains make up material cognitive systems. In: Gudwin, R., Queiroz, J. (eds.) Semiotics and Intelligent Systems Development. Idea Group Inc., Hershey (2007)
21. Haidt, J., Joseph, C.: Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Daedalus **133**, 55–66 (2004)
22. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**(4), 814 (2001)
23. Magnani, L.: Understanding Violence. Springer, Dordrecht (2011)
24. Batson, C.D.: What's wrong with morality? Emot. Rev. **3**, 230–236 (2011)
25. Sulmasy, D.: What is conscience and why is respect for it so important? Theor. Med. Bioeth. **29**, 135–149 (2008)
26. Barsalou, L.W.: Perceptual symbol systems. Behav. Brain Sci. **22**, 577–660 (1999)
27. Narvaez, D.: The embodied dynamism of moral becoming: reply to Haidt. Perspect. Psychol. Sci. **5**, 185–186 (2010)
28. Roeser, S.: Intuitions, emotions and gut reactions in decisions about risks: towards a different interpretation of 'neuroethics'. J. Risk Res. **13**, 175–190 (2010)
29. Kirkman, R.: Through the looking-glass: environmentalism and the problem of freedom. J. Value Inq. **36**(1), 29–43 (2002)
30. Haidt, J.: The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom. Basic Books, New York (2006)
31. Sun, R.: Duality of the Mind: A bottom-up approach toward cognition. Mahwah, N.J: L. Erlbaum Associates (2001)
32. Smith, A.: The theory of moral sentiments: Raphael, D.D., Macfie,A.L. (eds.) Glasgow Edition of the Works and Correspondence of Adam Smith, vol. I. Liberty Fund, Indianapolis (1982). http://oll.libertyfund.org/title/192
33. White, J.: Manufacturing morality, a general theory of moral agency grounding computational implementations: the ACTWith model. In: Floares, A. (ed.) Computational Intelligence. Nova Science Publishers, Hauppauge (2012)
34. White, J.: An information processing model of psychopathy and anti-social personality disorders integrating neural and psychological accounts towards the assay of social implications of psychopathic agents. In: Fruili, A.S., Veneto, L.D. (eds.) Psychology of Morality. Nova Science Publishers, Hauppauge (2012)
35. White, J.: Autonomy rebuilt: rethinking traditional ethics towards a comprehensive account of autonomous moral agency. Nat. Intell. **1**, 32–39 (2012)
36. White, J.: Conscience: toward the mechanism of morality. University of Missouri-Columbia (2006)

37. White, J.: Understanding and augmenting human morality, the ACTWith model. In: Magnani, L, Pizzi, C., Carnielli W. (eds.) Studies in Computational Intelligence #314: Model-Based Reasoning in Science and Technology, pp. 607–620. Springer, Heidelberg (2010)
38. Kant, I.: The Critique of Practical Reason, (trans. Abbott, T.K. 1788) Pennsylvania State University Electronic Classics Series (2010). http://www2.hn.psu.edu/faculty/jmanis/kant/Critique-Practical-Reason.pdf
39. King, M.L., Jr.: The other America. http://www.gphistorical.org/mlk/mlkspeech/index.htm (1968)
40. Alexander, T.: John Dewey and the moral imagination: beyond Putnam and Rorty toward a postmodern ethics. Trans. Charles S. Peirce Soc. **29**, 369–400 (1993)
41. Fesmire, S.: John Dewey and moral imagination: pragmatism in ethics. Indiana University Press, Bloomington (2003)
42. Young, L., Koenigs, M.: Investigating emotion in moral cognition: a review of evidence from functional neuroimaging and neuropsychology. Br. Med. Bull. **84**, 69–79 (2007)

Part III
Historical, Epistemological, and Technological Issues

Metaphor and Model-Based Reasoning in Maxwell's Mathematical Physics

Ryan D. Tweney

Abstract The role of model-based reasoning in experimental and theoretical scientific thinking has been extensively studied. However, little work has been done on the role of mathematical representations in such thinking. I will describe how the nature of mathematical expressions in physics can be analyzed using an extension of the metaphoric analysis of mathematics. Lakoff and Núñez [29] argued that embodied metaphors underlie basic mathematical ideas (e.g., the concept of "number" is based on the embodied operations of "collecting objects"), with more complex expressions developed via conceptual blends from simpler expressions (e.g., "addition" as "combining collections"). In physics, however, the need to represent physical processes and observed entities (including measurements) places different demands on the blending processes. In model-based reasoning, conceptual blends must often be based on immediately available embodiments as well as highly developed mathematical expressions that draw upon long term working memory. Thus, Faraday's representations of magnetic fields as "lines of force" were modeled by Maxwell as vectors. In the paper, I compare Faraday's experimental investigation of the magnetic field within a magnet to Maxwell's mathematical treatment of the same problem. Both can be understood by unpacking the metaphoric underpinnings as physical representations. The implications for analogical and model-based reasoning accounts of scientific thinking are discussed.

> Go to! Prepare your mental bricks!
>
> Fetch them from every quarter,
> Firm on the sand your basement fix
> With best sensation mortar.
> The top shall rise to heaven on high–
> Or such an elevation,

R. D. Tweney (✉)
Bowling Green State University, Bowling Green, OH 43403, USA
e-mail: tweney@bgsu.edu

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_21, © Springer-Verlag Berlin Heidelberg 2014

> That the swift whirl with which we fly
> Shall conquer gravitation.
> Maxwell ([34], p. 636).

Mathematics is central in science; it is frequently used as a basis for calculation, as a means of derivation of new expressions, and—the focus of the present paper—as a means of *representation*. Oddly, however, there are few attempts to deal with the power of mathematics as a representational medium in science, in spite of extensive work on the psychological underpinnings of scientific thought in general (see, e.g., [12]).

To clarify what I mean by representation, consider the following. In his *Principia mathematica* (1687), Isaac Newton formulated a law of universal gravitation which is usually today expressed with the following equation:

$$F = G m_1 m_2 \big/ r^2 \tag{1}$$

The equation gives the force, *F*, between two bodies of mass m_1 and m_2, separated by the distance *r*; *G* is the universal gravitational constant. About a century after Newton, Lagrange [28] showed that there was an alternate way to represent the dynamics among a system of bodies:

$$L = T - U. \tag{2}$$

Here, *L*, now known as the Lagrangian, is given as the difference between the kinetic energy, *T*, and the potential energy, *U*, quantities which can be defined for every point in the space between two (or more) objects. Lagrange showed that his formulation could solve all the same problems as Newton's and in many cases was easier to use; it possessed both calculation advantages and led to new derivational possibilities.

In fact, however, the two expressions, Newton's and Lagrange's, are fundamentally different in the way they represent the same physical reality. Newton's is based on an "action at a distance" view; it tells the relation between the masses in terms of the distance between them, but says nothing about the intervening space; the gravitational attraction just "happens" across empty space. Lagrange's equation, however, is defined at every location *between* the masses. Thus, where Newton's is nonlocal in character, Lagrange's equation is local. In this sense, it is more compatible with a field-like conception of the gravitational forces. As representations, therefore, the two expressions convey something entirely different about the dynamics of gravitational attractions. Furthermore, given this difference, it is appropriate to ask what effects the differing representations might have on the way in which physicists reason about gravitation. In particular, how might such representational differences affect the model-based reasoning of a physicist?

It is a commonplace to say that different kinds of mathematics are needed to deal with different kinds of physics. Thus, an Aristotelian world view, which is focused upon a world of objects, is associated with Euclidean geometry, an extremely powerful way of dealing with object shape and size. During the seventeenth century, and the emergence of analytic geometry, it became easier to talk about the relations among objects. For example, one could readily determine the intersection between two curved lines or surfaces. The physics that emerged as the result of the Galilean/Newtonian world view in turn drove the development of calculus as a means of determining and describing the motion of objects, and, more generally, of changing quantities in general. The eighteenth century saw extensive development of the tools of calculus (Lagrange's work being just one example), a development that continued in the nineteenth century [22].

But the nineteenth century brought a new kind of physics on the scene, one based on field theory, as exemplified by the work of Michael Faraday and James Clerk Maxwell. New ways of thinking required new kinds of mathematics, the mathematics of space-filling, vectorial, expressions. Field theories require attention to the entire space surrounding objects (and even, as we shall see, inside the objects) and gave new meaning to Lagrange's approach. For field theories, Euclidean formulations are inadequate, and even analytic geometric methods can be tedious and unilluminating. Developments in the calculus during the eighteenth century overcame these limitations to a large extent; in particular, as partial differential equations became available, it became easier to represent phenomena that were not tied to the object-centered world of objects and motion [15].

Note that in speaking of the representational power of mathematical expressions, I am really talking about the use of mathematics in science, rather than of mathematics as such. Both Newton and Maxwell, for example, were powerful mathematical thinkers, but they were also finely tuned to the representational use of mathematical expressions. For Newton, this centered on a geometric mathematics; for modern readers, his *Principia mathematica* (1687) is difficult to read (in part) because we have lost the feel for how his physics can be represented in this way. Translated into the (today) more familiar Leibnizean notation for the calculus, however, the underlying representations become more transparent. For Maxwell, the notation is more familiar (to those who have had a physics course in electricity and magnetism). While some translation is still needed [13], Maxwell's field-like use of integral and differential vector expressions, as exemplified in his *Treatise on electricity and magnetism* (1873/1891) is still important.

Newton's mechanics, especially as it was understood after Newton, assumed that the fundamental principle of motion depended upon forces that acted at a distance. Two masses attract each other because the gravitational force centered on each produces the motion. Throughout the eighteenth and most of the nineteenth centuries, similar action at a distance forces were presumed to be responsible for electric and magnetic actions. Just as gravitational force obeys an "inverse square law" (as in Eq. 1, above), so also did the attractive or repulsive force between two magnets or two electric circuits. The action at a distance account was challenged by Michael Faraday, who instead argued that electric and magnetic forces depended upon "lines of force"; the first true field theory in physics. By the end of his life, Faraday believed he had demonstrated the physical reality of the lines as immaterial but real centers of "power" [20].

Faraday was justly well-known for his many experimental researches and discoveries, but his theoretical account had almost no adherents—except the young Maxwell. For Maxwell, Faraday's account was a seminal one, and he set about to translate it into mathematical expressions. Eventually, he was able to show that the prevalent action at a distance theories of electromagnetic effects were less tenable than a true field theory (although this account also was slow to gain acceptance, as Hunt, [24], has shown).

In the present paper, using a part of Maxwell's account, I will attempt to show how cognitive science can provide an analytic framework for an understanding of the role of mathematics in physics. Maxwell's reformulation of classical physical ideas can thus be understood in cognitive terms, using recent formulations of model-based reasoning in science, and recent analyses of the underlying metaphoric bases of mathematics. The argument is based on three claims: (1) that mathematical representations can serve in model-based reasoning, and (2) that an understanding of how they are used requires attention to the embodied metaphoric understandings of the expressions. The metaphoric bases are in turn (3) dependent upon automated cognitive processes related to the employment of long term working memory. In this way, the external representation in the form of a mathematical expression is coordinated with an internal representation.

One terminological point is needed. In distinguishing between metaphors and analogies, I am using an unconventional division between two terms often seen as interchangeable. In the present usage (following [51]), I use *metaphor* to signal a taken-for-granted, tacit, comparison. I use *analogy* to signal a comparison between a source and a target that must be explicitly argued. In the particular case of Maxwell's physics, there have been many studies of his use of analogy in this sense, but little about his use of metaphor.

1 Cognitive Tools for Interpretive Understanding

Each of the three cognitive claims has a somewhat different epistemic grounding. The first claim I take to be given. That is, abundant research and scholarship, some reflected in the other papers in this volume, have shown that model based reasoning is ubiquitous in science—this will not be argued as such in the present paper. On the other hand, the embodied metaphoric claim is an extension of a current approach, one which is not without controversy. While I will not review the pros and cons, nor "claim sides," I do hope to convince the reader that use of a metaphoric analysis of the tacit, taken-for-granted, aspects of mathematical physics can illuminate the representational power of mathematics.[1] Finally, I use recent research on expertise and long term working memory as an explanatory

[1] The embodiment of metaphor will also be assumed here, and is important to the notion of model based reasoning as a species of abduction (e.g., [33]). For further discussion of these issues, see Cat [3], Gooding [21], Nersessian [39], and Tweney [51]. Simpson [42, 41], while emphasizing the rhetoric of Maxwell's *Treatise*, is advancing a similar argument.

tool, a way of justifying the metaphoric analysis and of suggesting ways in which model based reasoning can be learned and acquired as a working tool.

Model-Based Reasoning. Model-based reasoning rests on the claim that scientific thinking is largely a matter of the development of mental models of a richly varied sort; models that are involved in constructive and manipulative activities and that draw upon information in various formats, including linguistic and imagistic simulations, as well as external representations [39]. The traditional cognitive views of mental models (e.g., [26]), which centered on linguistic and propositional reasoning, have been extended in their application to scientific thinking. Thus, Nersessian [39], drawing partly on Maxwell's use of analogy, described a model-based reasoning process which included the mental simulation of complex physical systems (see also [14]). Clement [5] emphasized the recursive character of model based reasoning, arguing for a "Generate-Evaluate-Modify" cycle. As with Nersessian's approach, Clement emphasized the way in which scientific models are successively modified and tested. By studying both scientists and advanced college students in real time, Clement was able to track these processes from their initial formulations to the final, tested and justified, model.

Metaphoric Processes. In recent years, linguists and cognitive scientists have explored the metaphoric underpinnings of language. The claim is that common expressions like "falling in love," or "building an argument" are actually based on the specific metaphors of physical falling or of building construction. This has been argued as a way to connect the abstractness of language with sensorimotor cognition, and of *embodied* cognition in general [30].

Lakoff and Núñez [31] argued that even the most "abstract" of mathematical formulations are also grounded in basic cognitive embodiments via the use of metaphor. For example, the arithmetical operation of addition is related to the elementary cognitive operations of collecting objects. Thus "Object collection" as source is mapped onto "Arithmetic" as target. "Collections of objects of the same size" are mapped onto "Numbers," "Putting collections together" onto "Addition," and "Taking a smaller collection from a larger one" onto "Subtraction" [31]. Arithmetic itself can then become the source for further extensions to new target domains. "Grounding metaphors" according to Lakoff and Núñez are linked directly to sensorimotor experience (as in the examples), and these are then the source for further conceptual metaphors.

Turner [48] has argued that "conceptual integration," the "blending" of disparate conceptual spaces is a basic cognitive operation that underlies the emergence of new meaning. Thus, in the metaphor, "The surgeon is a butcher," the spaces corresponding to the source and target of the metaphor each contribute some meanings to the blend, but the emergent meaning of the whole is something not characteristic of either of the "parent" spaces. Turner has shown how non-Euclidean geometry can be interpreted as a conceptual blending from Euclidean geometry (see [48], esp. Appendix C, pp. 163–168). In this fashion, as Lakoff and Núñez also argue, the seemingly abstract spaces of mathematics can be unwrapped by showing their origins in successively more basic conceptual spaces. The approach is general; for example, Núñez [40] has used it to interpret the historical case of the development of transfinite cardinal numbers by Georg Cantor.

The conceptual theory of metaphor and its role in science has been the subject of some controversy (see, e.g., the critiques by [36, 35, 50], and the replies by [18] and by [19]). Still, for present purposes, in which the approach is used to structure an interpretive framework, the outcome of the controversy is not directly relevant. For the present analysis, what counts is the ability of the approach to provide a tool for the untangling of what is usually implicit in mathematical physics.[2]

Long Term Working Memory. Cognitive scientists have long distinguished between (1) short term memory (STM), which holds a limited amount of new information for a brief time, (2) long term memory (LTM), a larger, more permanent, store, and (3) working memory (WM) which holds material recently retrieved from LTM as needed in a specific task. Ericsson and Kintsch [7] extended the concept of WM by noting that, among experts, specific kinds of processes seemed to be taking place when domain-specific material was retrieved. Referring to this as (4) long term *working* memory (LTWM), Ericsson and Kintsch suggested that many of the results of expertise can be explained by the emergence of LTWM. In particular, rather than relying upon specific retrieval cues, experts have acquired *structured* retrieval mechanisms to bring domain-relevant material and skills into working memory. In the case of mathematical reasoning in science, a differential equation, say, can be thought of as entraining a series of other components of the knowledge of calculus into LTWM.

Ericsson and Kintsch showed that expert readers (but not inexpert readers) can keep the thread of a book's argument "in mind" long after the contents of ordinary (short term) working memory have been replaced by new information. In effect, long term memory remains immediately accessible. Experts thus have a specific set of retrieval structures that make this possible. The relevant skill is more than simply possession of a set of retrieval cues. The expert retrieval structures also imply an anticipatory element that flags what might be relevant in the near or far distant future. Such structures are domain-specific and develop only after extensive deliberate practice. In the case of expert reading, the larger gist of text remains available across long stretches of text. The same is true for differential equations in physics [27].

For Ericsson and Kintsch, LTWM is acquired as an aspect of the acquisition of expertise and comes about via the extended deliberate practice characteristic of the highest levels of expertise. Thus, physics professors perform differently than physics graduate students on problems where both have the same specific content knowledge, as Chi et al. [4] have shown (see also [32]). The professors have developed such LTWM retrieval structures centered around the basic principles of

[2] Note also that my approach differs from accounts that regard metaphor as a somewhat loose use of similarity, while analogy has been regarded as founded on more severe constraints. See Gentner and Jerzioski [17], which adopts such a view. I am using the two terms in unconventional fashion, with metaphor referring to implicit comparisons and analogy to those drawn explicitly.

physics, while the graduate students are still acquiring them and are more dependent upon surface-level cues.

2 Maxwell's Use of Mathematical Representation

In the previous section, three cognitive concepts have been outlined; these will serve as the interpretive framework for the following discussion. I will argue (a) that the mathematical representations used in physics exemplify model based reasoning, (b) that the functioning of such models depends upon acquired metaphors and conceptual blends, and (c) that the acquisition of such metaphoric foundations can be explained by the development of long term working memory. To illustrate the argument, I will develop an analysis of one part of Maxwell's field theory of electromagnetism, a "mini" case study. To provide context, I give a brief account of experimental work by Faraday which is directly relevant.

From Faraday to Maxwell. In a conventional view that finds its way into many textbooks, Michael Faraday (1791–1867) was one of the greatest experimental scientists of the nineteenth century, responsible for a long string of discoveries, most famously in electricity and magnetism. Still, his theoretical ideas were couched in a non-mathematical language that did not, by and large, appeal to his contemporaries. By contrast, as the conventional view has it, James Clerk Maxwell (1831–1879), was one of the greatest mathematical physicists of the century. His "translation" of Faraday's theory into mathematical expressions and his subsequent extension of those theories was the ultimate triumph of classical physics.

The conventional view, while broadly correct, misses the nuances of the relation between Faraday's theory and Maxwell's.[3] In particular, Maxwell saw in Faraday an intuitive mathematician of the highest order: "As I proceeded with the study of Faraday, I perceived that his method of conceiving the phenomena was also a mathematical one, though not exhibited in the conventional form of mathematical symbols" ([35], Vol. 1, p. ix). In the case described below, this will become more clear.

Across thousands of experiments, Faraday developed by 1850 a coherent theory of electric and magnetic fields and the relation between the two [11, 20, 49]. Centering on the notion of "lines of force," which he conceptualized as space-filling immaterial entities possessing dynamic properties, he argued that these were physically real and that his experiments had proved their existence and determined many of their properties. Faraday acknowledged the incompleteness of the theory, in part because it was not possible to determine the velocity with which such fields moved. And, while he had shown by experiment [9] a possible relation between electromagnetic fields and light (in the form of a rotation of the direction of

[3] A good brief introduction to Faraday's work is James [25]. For Maxwell, a good beginning is Everitt [8]. Both works acknowledge the nuances!

polarization of light when traversing a dense transparent glass subjected to a strong magnetic field), he could only speculate on the physical nature of the relationship.

Thomson [45, 46] was the first to attempt a mathematical treatment of Faraday's lines of force. Thomson showed that there was an analogy between the equations describing the distribution of electric and magnetic force and the equations describing the distribution of heat within a solid. In developing the analogy, Thomson took no position on the reality of the lines of force, although he later claimed that the equations constituted "a full theory of the characteristics of the lines of force" (Thomson [45], p. 1, footnote added in 1854).

Maxwell began his account of Faraday's theory in a series of three papers in 1855–1856, 1861–1862, and 1864, and summarized the final state of his theory in the 1873 *Treatise on electricity and magnetism* [35]. The development across the three early papers has been extensively analyzed (see especially [38, 39, 41]). In the course of the three papers, Maxwell did in fact "translate" Faraday's theory into mathematical form (as the conventional view has it), but there were significant changes along the way. Beginning, like Thomson, with an analogy, Maxwell considered the lines of force in Faraday's theory as if they were tubes carrying an incompressible fluid, then developed a mechanical model (based on vortex wheels in a mechanical ether, again, as an analogy), and finally re-expressed Faraday's notion of *force* into a new form, one based on a dynamical theory with *energy* as the focus; this last view was then fully developed in the *Treatise*.[4]

Maxwell's *Treatise* is a complex work with multiple goals. Conceived as a textbook, it includes much material on the fundamental empirical facts of electricity and magnetism, accounts of experiments and measuring devices, and a "dialectical" development of the final theory (see [13, 42, 43]). In its modern form, Maxwell's final account is summarized as "Maxwell's Equations," four vector equations that represent the electric and magnetic fields and the relation between the two. In a previous paper, I have outlined how the four equations can be tied metaphorically to primitive embodied notions of stress and strain [51]. Here, I compare one aspect of Maxwell's treatment of magnetism to a parallel case examined experimentally by Faraday. This, in turn, will allow an account of how the three cognitive schemas outlined in the previous section can be understood as the bases of mathematical representations in physics.

Faraday: Magnetic Lines Within a Magnet. The year 1846 was a crucial one for the development of Faraday's theory of magnetism. In that year, he published three papers on the nature of magnetic interactions, first with light [9], then with matter (a brief account is in [49] and a more thorough account in [20]). Confirming his belief that magnetic lines of force extended through all of space, even penetrating into material bodies, he argued that lines of force were "conducted" within the substance of material bodies, thus establishing that diamagnetic substances (such

[4] Maxwell's famous derivation, suggesting that light was an electromagnetic manifestation, appeared initially in the second paper, was re-expressed in the third paper, and finalized at the end of the *Treatise*.

as bismuth or glass), as well as paramagnetic substances (such as iron) were subject to magnetic influence. Further, by showing the rotation of a polarized light beam in a magnetic field, he was able to argue that magnetic lines of force were perhaps implicated in the nature of light.

Still, he needed to show that the lines of magnetic force could be observed even within the substance of a magnet and that they were closed curves [10]. To do this, he conducted an interesting series of experiments in which two long bar magnets of equal strength were placed side by side, with a small gap between (thus acting as a single, thicker, magnet). These were mounted on a shaft within an apparatus that allowed their rotation (Fig. 1). With commutators on the shaft of the rotating apparatus, he could then run a wire alongside or within the slot between the magnets, and the wire could be rotated together with or independently of the magnets (which were equivalent to a single magnet with a slot down the middle). Connecting the ends of the wire to a galvanometer, he was able to detect any induced currents in the wire.

Faraday tried a variety of configurations (I simplify his many arrangements in this description), first rotating the wire and magnet together (no current was produced), then the wire alone or the magnet alone (as in Fig. 1b, again no current was produced). He then separated the wire at the point b (Fig. 1c), which permitted the segments to be rotated separately while maintaining electrical contact. He found that rotating the magnet with the segment b–d and without the segment b–c *did* produce a current, and rotating the wire segment c–b without the segment b–d *also* produced a current, but in the opposite direction. He argued that the segment of the wire from b to d was cutting all the lines of force when revolved, as did segment a to c. Further, the size of the current produced was the same in both cases. Because the currents were in opposite directions, when the whole wire (a–d–b–c) was revolved and the magnet kept stationery, no current was observed: Each wire was cutting *all* the lines of force but the generated currents were in opposite directions, thus cancelling each other. This was the result he was after: the lines of magnetic force ran through the magnet, out at one end, curved around through space, and re-entered at the other end of the magnet. Magnetic lines of force are closed curves.

Maxwell: Magnetic Lines Within a Magnet. Maxwell's *Treatise* is divided into four parts, with the fourth part developing the final form of his theory of electromagnetism and the third presenting his account of magnetism. The first chapter of the third part considered the magnetic potential at any point outside of a nearby magnet, showing that the force on a "unit magnetic pole" is equal to the gradient of the potential (∇V, where V is a scalar function), i.e., to the rate of change of the potential in the direction of greatest change. In Chap. 2, Maxwell considered the forces within a magnet. In contrast to Faraday, however, he did not here conduct experiments, nor replicate Faraday's (although they are cited). Instead, he conducted a series of thought experiments.

He began by imagining a cylindrical hollow cavity within a bar magnet (Fig. 2). Taking its length as $2b$ and its radius as a, he then imagined a unit magnetic pole centered within the cavity. Such a pole is an imaginary object, since magnetic

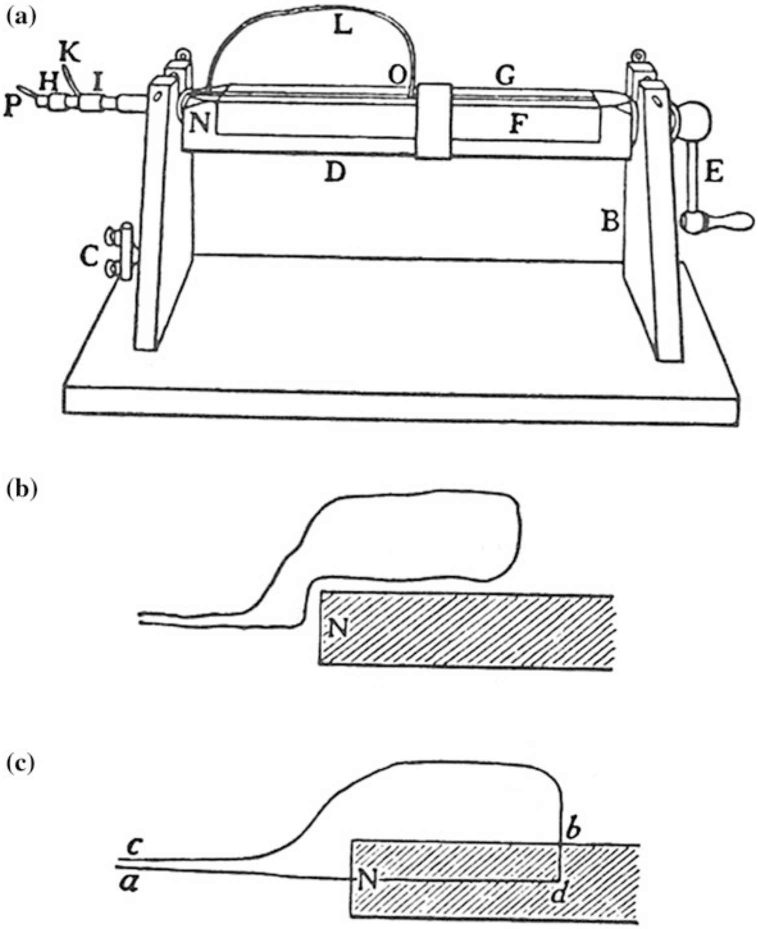


Fig. 1 Faraday's experiments on the lines of force within a magnet. **a** The apparatus used; *F* and *G* are two identical magnets mounted on a shaft with a small gap between. Commutators are shown at *H*, *I*, and *O*. **b** The wire separated from and entirely outside the magnet. **c** The wire run through the inside of the magnet. Segments *a–d* and *b–d* can be rotated independently, or together with *b–c*. No current is produced by the entire loop, or by *a–d* alone, but current is produced by *b–d* alone or by *c–b* alone (in the opposite direction ([10], pp. 333, 338)

monopoles do not exist (that is, if you break a magnet into two pieces, each piece will have a north and a south pole, breaking them again, each piece will have two poles, and so on). Still, *were such a thing to exist*, it is possible to represent the forces it would experience. There are two sources; first the forces due to magnetic induction from the ends of the cavity. Since the field lines are parallel to the walls of the cylinder, the walls play no role, only the circular ends are involved. Second, there are forces due to the potential field within the cavity. That is, there is an overall field because the cavity is within a magnet and a specific field due to the

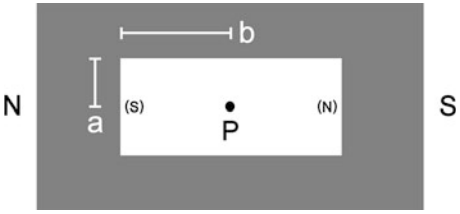


Fig. 2 Maxwell's thought experiment: a bar magnet with a cavity inside. The cavity is cylindrical, of length 2 *b* and with faces of radius *a*. Note that the polarity of the faces is the reverse of the polarity of the nearest end of the magnet

surface distribution of magnetism on the ends of the cylindrical cavity. Note that the forces due to the circular surfaces are of opposite polarity to the ends of the magnet.

Maxwell first considered the field due to the surface distribution on the cylinder ends, claiming that the forces on the monopole are equal and in the same direction (because the monopole will be attracted by one surface and repelled by the other). This force will be:

$$R = 4\pi I\left(1 - \frac{b}{\sqrt{a^2 + b^2}}\right) \qquad (3)$$

where R is the force and I is the intensity of magnetization. Because the dimensions of the cavity are involved, the force is dependent upon the shape of the cavity. Interestingly, Maxwell does not show how this equation is obtained, taking for granted that the reader will know how to do this (while not lengthy, I will not carry this out—see [13], comment on 396.2, and the discussion below).

With this in hand, Maxwell now asked the reader to consider two cases. In the first, imagine that a is very small, that is, shrink the diameter of the cylinder cavity. From Eq. (3), note that R will approach 0 as a approaches 0. In the second case, let the cylinder shrink in length. As b approaches 0, then R approaches $4\pi I$. This means that, in the first case, a long and thin cylinder, the force will simply be that due to the overall field; it will be the gradient of the potential. Maxwell calls this "magnetic force within the magnet" and symbolized it as a vector, $\boldsymbol{H}$ (here using bold-face, to indicate a vector). In the second case, which becomes a flat disk as the cylinder length shrinks, the force is dependent on R and is compounded of $4\pi I$ and $\boldsymbol{H}$. He symbolizes this new quantity as $\boldsymbol{B}$ and calls it the "magnetic induction." The two terms are related by a simple equation, via the overall intensity of magnetization, I, which, written as a vector, is $\boldsymbol{I}$:

$$\boldsymbol{B} = \boldsymbol{H} + 4\pi\boldsymbol{I}. \qquad (4)$$

Note from Eq. (4) that the distinction between $\boldsymbol{B}$ and $\boldsymbol{H}$ will hold only within a magnet; in the absence of a surrounding magnet, that is, when $\boldsymbol{I} = 0$, the two are identical (Fig. 3).

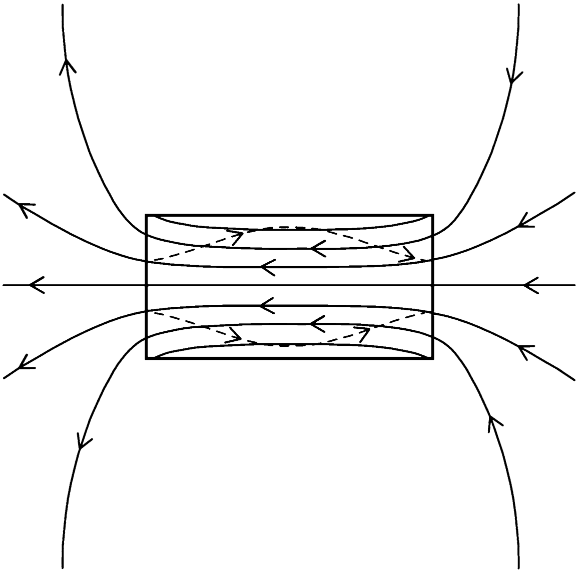


Fig. 3 Showing the lines of force within and without a bar magnet; the north pole is to the left. *Solid lines* are lines of induction, ***B***; *dashed lines* are lines of ***H***. On the outside of the magnet, the two fields are identical

Maxwell used the relation between ***B*** and ***H*** to clarify a paradox in Faraday's notion of lines of force. The paradox arose because the directions of ***B*** and of ***H*** differ. That is, the magnetic force due to ***H*** always goes from the north pole of the magnet to the south pole—both inside and outside the magnet! As a result, they meet "head on," as it were, at the south pole, which then constitutes a termination. But Faraday's lines of force are continuous closed curves and do not meet; they run from the north pole to the south pole when outside the magnet and continue from the south pole to the north pole inside the magnet, thus constituting closed curves. The magnetic lines due to ***B*** have the needed property—they run from north to south outside and from south to north inside. For this reason, "All that Faraday says about lines of force… is mathematically true, if understood of the lines… of magnetic induction" ([35], Vol. 2, p. 28).[5]

Maxwell's clarification of the difference between ***B*** and ***H*** was highly consequential. In particular, it allowed him to distinguish between magnetic effects which were mechanical in nature and those which were involved in the production of currents in a nearby conductor. In the case of ***H***, one is speaking of the "magnetic force," and this is purely mechanical and can be manifested by the effect on a compass needle or an iron filing. In the case of ***B***, the "magnetic induction," the force can be manifested as an *electromotive* force, that is, as one

[5] Thomson [47] had considered a problem similar to Maxwell's, in determining the force on a unit pole placed within a small cavity in a magnet. However he did not resolve the directional paradox between the directions of what were later called ***B*** and ***H*** by Maxwell. Smith and Wise ([44], pp. 279–281) describe Thomson's approach and indicate that he did not fully publish his results.

producing a current in a conductor. For Maxwell (mathematically), as for Faraday (experimentally), these corresponded to two different ways to detect the presence of a magnetic field.

Faraday initially mapped magnetic fields by using a small magnetized needle suspended from a thread; such a needle will orient itself along lines of force. He later used a small loop of wire attached to a sensitive galvanometer. When moved in a field, a current would be generated in the loop, a current detectable by a sensitive galvanometer at a distance. This "moving wire" became his favored method, mapping, in Maxwell's terms, the lines of induction. In subsequent chapters of the *Treatise*, Maxwell developed the mathematical representation of such mappings in great detail, arguing that the magnetic induction and not the magnetic force is the physically significant quantity.

3 Unpacking the Model-Based Reasoning

How do we use the cognitive framework sketched in the first part of this paper to understand the case study? Maxwell, like Faraday, used model based reasoning in the example, as should be clear. Like Faraday, Maxwell described for the reader a series of actively used constructions to make the argument for forces within a magnet. Faraday described actual experiments, inviting the reader to construct a mental model of the apparatus, procedures, and results. Maxwell used a thought experiment in the same way; that is, his reader was asked to construct a mental model of an "experiment" that could be done only in the mind's eye and not in reality. That is not to say that actual experiments were entirely absent in Maxwell's account, rather, they were presumed to be present in the reader's knowledge, based, in part, on the previous chapters of the book and in the references to Faraday's experiments.

Still, it is not the case that we can make a one-to-one mapping between the kinds of knowledge that underlies Faraday's reasoning and that of Maxwell. This is because Maxwell also had to rely upon a kind of knowledge base not used by Faraday. In particular, Maxwell relied upon the *metaphoric* understandings embedded within the mathematical expressions used. For example, consider Eq. (3) from the previous section:

$$R = 4\pi I\left(1 - \frac{b}{\sqrt{a^2 + b^2}}\right) \tag{5}$$

I noted above that Maxwell did not provide a derivation of this result, instead assuming that his readers would be able to recognize it. To show its metaphoric nature, first consider the term $a^2 + b^2$. From Fig. 2, it is apparent that this is related, via the Pythagorean Theorem, to the length of the hypotenuse of the triangle with sides a and b. If we take the square root and call this r, then we can simplify Eq. (5):

$$R = 4\pi I\left(1 - \frac{b}{r}\right) \tag{5a}$$

This, in turn, becomes

$$R = 4\pi I \ - \ 4\pi I\frac{b}{r} \tag{5b}$$

Suppose now that *a* shrinks (Maxwell's first case). Then *b/r* goes to one and *R* goes to zero. And if *b* shrinks (Maxwell's second case), then *R* goes to $4\pi I$.

The attentive reader can now see how the metaphoric underpinnings worked in the discussion of this equation. For, in fact, what I have asked *you* to do is what Maxwell (with, to be sure, more extensive metaphors assumed) asked of his readers! That is, I drew upon your knowledge of the Pythagorean Theorem and upon your metaphoric sense of what happens when geometric terms like *a* and *b* change. Further, your sense of how algebraic equations can be modified, as in going from Eq. (5) to (5a) and (5b), was also involved. These did not need to be specifically argued because, as Lakoff and Núñez [31] argued, these have been acquired on the basis of long practice—they are conceptual blends with metaphoric groundings. On my account, they are not analogies, because the links between source and target are implicit and assumed to be known among his readers. This is why Maxwell does not explicate Eq. (5).

However, Eq. (5) is not yet fully explicated for our purposes. Where does the $4\pi I$ come from? In the previous chapter, Maxwell had considered the force on a small magnet due to the distribution of a surface of magnetic "matter" (like the imagined cavity and the magnetic monopole, this is another convenient fiction). That discussion, in turn, relied upon results achieved in the first volume of the *Treatise*, in which he showed that the surface distribution of an electric charge on a conductor exerted a force near to the conductor equal to $4\pi\sigma$, where σ is the surface distribution of charge. In the present case, *I* is equivalent to the charge in the earlier case. In particular, both charge and magnetic entities exert force according to an inverse square law, that is, inversely as the square of the distance. Thus, $4\pi I$, unlike the other part of Eq. (3) *is* an analogy, albeit itself grounded in the mathematics of earlier parts of the book: "Since the expression of the law of force between given quantities of 'Magnetism' has exactly the same mathematical form as the law of force between quantities of 'Electricity' of equal numerical value, much of the mathematical treatment of magnetism must be similar to that of electricity" ([35], Vol. 2, p. 5). Maxwell is able to carry over the expression for the magnetic surface density from the equivalent expression for electric surface density: he does not need to repeat the derivation (which is also built on metaphoric grounds and hence can be taken as given), he only needs to have shown the analogy.

We can again obtain an informal understanding by asking where the multiplier 4π comes from. Note first that the monopole at point *P* in Fig. 2 is subjected to an attractive force from one face of the cylindrical cavity and a repelling force from the other face. Both forces are in the same direction, so any one face is contributing 2π to the result. But 2π is the circumference of a circle of radius 1. Here, it appears as if Maxwell is relying upon a previous result from the first volume of the *Treatise*, namely Stokes's Theorem, which states that the surface integral of a function describing a surface is equal to the line integral of the curve bounding that surface. Explaining this would go beyond the scope of the present paper, but it implies in the case of the circular face of the cavity that the force due to the face can be construed as either based on the density of magnetization of the surface or, equivalently, as based on a circulation around the closed curve (the circle) that bounds it. Thus, 2π emerges!

Note that for Maxwell's readers Stokes's Theorem would have been assumed knowledge (it is explained in a "Preliminary" chapter, [35], Vol. 1, p. 29). For the present purpose, however, it is enough to catch some glimpse of how the factor emerges; in the following chapter, Maxwell uses Stokes's Theorem to make a more explicitly physical representation. There, he shows that a magnetic "shell" (a surface bounded by a closed curve) can be represented equivalently by an electric current in a conductor that follows the same closed curve.

One final point: Maxwell's *Treatise* is notable in part for its use of vectors as representational entities. In the selection here, these appear as **H**, **B**, and **I**. I've previously discussed the metaphoric basis of vector representations [51]. For the present case, it needs only to be noted that vectors are quantities that represent both magnitude *and* direction. They can be grounded on elementary notions of muscular force and direction, and can then be conceptually blended with other mathematical concepts. Throughout the *Treatise*, Maxwell uses them (and the vector calculus) as part of his overall representation of fields (as in Fig. 3). The introduction of such vector analysis was an important milestone in mathematical physics generally, one that continues to be used today [6, 15].

4 Cognition and Metaphor in Mathematical Physics

The present paper has presented a sketch of a mode of analysis that has important implications for understanding how mathematical representations have gained such great importance in science. There have been many analyses of the role of analogy in model based reasoning, even extending to accounts of Maxwell's physics. However, the metaphoric aspect of mathematical representations holds the key to understanding how the tacit knowledge embedded within mathematical expressions can become an active part of model based reasoning.

Three points were made in the Introduction to this paper, that mathematical physics does involve model based reasoning, that metaphor underlies the representational use of mathematics, and that such metaphoric grounding is tacit and acquired (via LTWM) through the acquisition of expertise. I will discuss each in reverse order.

Since I have previously discussed the role of LTWM in Maxwell's case [50] only brief comment is needed here. Maxwell wrote the *Treatise* partly intending it as a text for the new Tripos exam in Natural Philosophy at Cambridge University (he had been appointed to the newly established professorship of natural philosophy in 1871; see [23, 52]). Maxwell's own education in science and mathematics (primarily at Edinburgh and Cambridge, but beginning even in his childhood) provided him with an extensive knowledge of the mathematics and physics that he took for granted in the book, and it is likely that he expected his students would have similar knowledge. He was writing the *Treatise* for those with the kind of retrieval structures that are fundamental to expertise. In recasting the case study for my readers, I have also made some assumptions; for example, that the reader would know the formula for the circumference of a circle, have algebraic skills, and know at least something about electricity and magnetism. Access to all of these relies upon a similar LTWM capacity; the cognitive underpinnings for Maxwell's students and my readers were not different in principle.

I have also previously spelled out the role of metaphors in the understanding of mathematical physics, using the modern form of Maxwell's Equations [51]. For the present case study, I have instead relied more closely upon the actual text written by Maxwell. Although closer to Maxwell's argument, much has been left out. Further, the analysis is informal and adapted to my readership, not Maxwell's. In this sense, what I have provided is not an analysis of the actual historical materials, but rather a reconstruction of a "possible world."[6] Even so, it should be possible to see the way in which metaphors and conceptual blends play a role in the arguments made by Maxwell. The analytic task here is to work backwards from the argument as presented by Maxwell to the underlying structure of the mathematical representations.

While a great deal has been written about Maxwell's use of analogy (see especially [39, 41]), I believe my analyses are the first attempts to use metaphor and conceptual blends to describe the *tacit* knowledge which Maxwell brought to bear on his arguments (see, for a similar attempt, not rooted in metaphor in the same fashion, [2]). It has been argued that, from a cognitive point of view, there is no inherent difference between analogy and metaphor.[7] The best-known such argument is due to Gentner (e.g., [16]). Her "structure mapping theory" of analogy and metaphor is based on the processes involved in mapping relations from a source to a target, and there is much evidence to suggest that this correctly captures many of the phenomena. Nersessian has extended this, describing (using Maxwell, in part) how analogies can participate in the creation of new conceptual content in science. In turn, the present paper supplements and extends all of these accounts.

[6] It is interesting to note the similarity of this maneuver to that used by Lakatos in *Proofs and refutations*, which used a similar ploy to discuss the nature of discovery in mathematical proof [29].

[7] Indeed, the terms, analogy and metaphor, have had a flexible boundary in much of the writing about their use in science. Thus, for example, much of what Bradie [1] has written about metaphor applies equally to analogy.

Model based reasoning is an integral part of many naturalistic accounts of science, and the present paper is no exception. That Maxwell used it in presenting his analysis of the magnetic field within a magnet should be clear, even from the brief segment considered. The thought experiment he presents is fundamentally anchored in the reader's ability to follow the claims made via the construction of a model and via the implementation of the mathematical representations involved. Note that they lead up to the expression of an *identity*, not an *equation* in the usual sense. That is, Eq. (4), $\boldsymbol{B} = \boldsymbol{H} + 4\pi\boldsymbol{I}$, is presented, not because it has calculational uses but because it shows the reader the relationships among key terms and because, by using vector notation for the first time in the section, it reiterates the directional character of the lines of induction, of force, and of magnetic intensity. It is important because of its representational character.

I noted earlier that Faraday represented magnetic "lines of force" experimentally, by constructing apparatus that enabled the detection of the lines within a magnet. Maxwell achieved the same thing using a thought experiment, a move that allowed him to distinguish between $\boldsymbol{H}$ and $\boldsymbol{B}$, thus identifying $\boldsymbol{B}$ as the physically significant quantity. The two approaches complement each other in an interesting fashion. Thus, Faraday's science is replete with "hand-eye-mind" representations; for him, the lines of force were physically real to the extent that he could observe their effects and manipulate their character. For Maxwell, the observation and manipulation were based, not on experiment directly, but on the expression of a mental model and its extension via the metaphoric underpinnings of the mathematical representations. It, too, had a "hand-eye-mind" character.

Ultimately, then, this is the true fashion in which Faraday and Maxwell can be seen as similar: both were doing science in a style dependent upon a fundamental *embodiment* of the conceptual representations they created. For both, this was, in fact, a conscious goal. When Faraday is seeking the physical reality of his lines of force, he is doing just what Maxwell was doing in seeking to identify the vector $\boldsymbol{B}$ as the "physically significant" quantity. That they followed different pathways, that Faraday's was experimental and Maxwell's mathematical, is not, in the end, the most important aspect for an understanding of their creative achievements. Instead, for both, the "mental bricks" of embodied representations were the means: "Firm on the sand your basement fix/With best sensation mortar."

Acknowledgments Thanks are due especially to Howard Fisher, who has saved me from many errors and is not responsible for remaining ones! I have benefitted greatly from discussions of Maxwell with John Clement, Howard Fisher, Frank James, Nancy Nersessian, and Thomas Simpson. The paper's ultimate origin stems from discussions with the late David Gooding and with Elke Kurz-Milcke. The proximate origin is a paper given at MBR012 in Sestri Levante, Italy, in June, 2012; I am grateful for the questions and comments of the other participants and to Lorenzo Magnani for his support. Matt Lira and Frank James provided helpful comments on an early draft, for which I am grateful.

References

1. Bradie, M.: Models and metaphors in science: the metaphorical turn. Protosociology **12**, 305–318 (1998)
2. Cat, J.: On understanding: Maxwell on the methods of illustration and scientific metaphor. Stud. Hist. Philos. Mod. Phys. **32**, 395–441 (2001)
3. Cat, J.: Into the 'regions of physical and metaphysical chaos': Maxwell's scientific metaphysics and natural philosophy of action (agency, determinacy and necessity from theology, moral philosophy and history to mathematics, theory and experiment). Stud. Hist. Philos. Sci. **43**, 91–104 (2011)
4. Chi, M.T.H., Feltovich, P.J., Glaser, R.: Categorization and representation of physics problems by experts and novices. Cogn. Sci. **5**, 121–152 (1981)
5. Clement, J.: Creative Model Construction in Scientists and Students: Imagery, Analogy, and Mental Simulation. Springer, Dordrecht (2008)
6. Crowe, M.J.: A History of Vector Analysis. University of Notre Dame Press, South Bend (1967)
7. Ericsson, K.A., Kintsch, W.: Long-term working memory. Psychol. Rev. **102**, 211–245 (1995)
8. Everitt, C.W.F.: James Clerk Maxwell: Physicist and Natural Philosopher. Charles Scribner's Sons, New York (1975)
9. Faraday, M.: Experimental Researches in Electricity, Nineteenth Series. On the Magnetization of Light and the Illumination of Magnetic Lines of Force. Taylor & Francis, London (1846) (Reprinted in M. Faraday (ed.) (1855) *Experimental researches in electricity* vol. 3, pp. 1–26)
10. Faraday, M.: Experimental Researches in Electricity, Twenty-Eighth series. On Lines of Magnetic Force: Their Definite Character; and Their Distribution Within a Magnet and Through Space. Taylor & Francis, London (1851) (Reprinted in M. Faraday (ed.) (1855) Experimental researches in electricity, vol. 3, pp. 328–370)
11. Faraday, M.: On the Physical Character of the Lines of Magnetic Force. Taylor & Francis, London (1852) (Reprinted in M. Faraday (ed.) (1855) Experimental researches in electricity, vol. 3, pp. 407–437)
12. Feist, G.J., Gorman, M.E. (eds.): Handbook of the Psychology of Science. Springer, New York (2013)
13. Fisher, H.: Maxwell's Treatise on Electricity and Magnetism: The Central Argument. Green Lion Press, Santa Fe (in press)
14. Forbus, K.: Reasoning about space and motion. In: Gentner, D., Stevens, A. (eds.) Mental Models, pp. 53–74. Lawrence Erlbaum, Hillsdale (1983)
15. Garber, E.: The Language of Physics: The Calculus and the Development of Theoretical Physics in Europe, 1750–1914. Birkhäuser, Boston (1999)
16. Gentner, D., Bowdle, B.: Metaphor as structure-mapping. In: Gibbs Jr, R.W. (ed.) The Cambridge Handbook of Metaphor and Thought, pp. 109–128. Cambridge University Press, Cambridge (2008)
17. Gentner, D., Jeziorski, : The shift from metaphor to analogy in Western science. In: Ortony, A. (ed.) Metaphor and Thought, 2nd edn, pp. 447–480. Cambridge University Press, Cambridge (1993)
18. Gibbs Jr, R.W.: Why many concepts are metaphorical. Cognition **61**, 309–319 (1996)
19. Gibbs Jr, R.W., Perlman, M.: Language understanding is grounded in experiential simulations: a response to Weiskopf. Stud. Hist. Philos. Sci. **41**, 305–308 (2010)
20. Gooding, D.C.: Final steps to the field theory: Faraday's study of magnetic phenomena. Hist. Stud. Phys. Sci. **11**, 231–275 (1981)
21. Gooding, D.: Experiment and the Making of Meaning: Human Agency in Scientific Observation and Experiment. Kluwer Academic Publishers, Dordrecht (1990)

22. Grattan-Guinness, I.: The Fontana History of the Mathematical Sciences: The Rainbow of Mathematics. Fontana Press, London (1997)
23. Harman, P.M.: The Natural Philosophy of James Clerk Maxwell. Cambridge University Press, Cambridge (1998)
24. Hunt, B.R.: The Maxwellians. Cornell University Press, Ithaca (2005)
25. James, F.A.J.L.: Michael Faraday: A Very Short Introduction. Oxford University Press, Oxford (2010)
26. Johnson-Laird, P.N.: Mental models in cognitive science. Cognitive Science 4, 71–115 (1980)
27. Kurz-Milcke, E.: The authority of representations. In: Kurz-Milcke, E., Gigerenzer, G. (eds.) Experts in Science and Society, pp. 281–302. Kluwer Academic/Plenum, New York (2004)
28. Lagrange, J.L.: Analytical Mechanics (trans. ed. by A.C. Boissonnade, and V.N. Vagliente). Kluwer Academic Publishers, Boston (1788/1997)
29. Lakatos, I.: Proofs and Refutations. Cambridge University Press, Cambridge (1963–1964/1976). (Originally published in British Journal for the Philosophy of Science, 14)
30. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and its Challenge to Modern Thought. Basic Books, New York (1999)
31. Lakoff, G., Núñez, R.E.: Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being. Basic Books, New York (2000)
32. Larkin, J.H., McDermott, J., Simon, D.P., Simon, H.A.: Models of competence in solving physics problems. Cogn. Sci. **4**, 317–345 (1980)
33. Magnani, L.: Abduction, Reason, and Science: Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
34. Maxwell, J.C. (undated, c. 1870/1882). To the chief musician upon Nabla. In: Campbell, L., Garnett, W. (eds.) (1882). The Life of James Clerk Maxwell, pp. 634–636. Macmillan, London
35. Maxwell, J.C.: A treatise on Electricity and Magnetism (2 volumes). Clarendon Press, Oxford. (1873/1891) (Third edition, revised by J.J. Thompson)
36. Murphy, G.L.: On metaphoric representation. Cognition **60**, 173–204 (1996)
37. Murphy, G.L.: Reasons to doubt the present evidence for metaphoric representation. Cognition **62**, 99–108 (1997)
38. Nersessian, N.: Faraday to Einstein: Constructing Meaning in Scientific Theories. Nijhoff, Dordrecht (1984)
39. Nersessian, N.J.: Creating Scientific Concepts. MIT Press, Cambridge (2008)
40. Núñez, R.E.: Creating mathematical infinites: metaphor, blending, and the beauty of transfinite cardinals. J. Pragmat. **37**, 1717–1741 (2005)
41. Siegel, D.M.: Innovation in Maxwell's Electromagnetic Theory: Molecular Vortices, Displacement Current, and Light. Cambridge University Press, Cambridge (1991)
42. Simpson, T.K.: Figures of Thought: A Literary Appreciation of Maxwell's Treatise on Electricity and Magnetism. Green Lion Press, Santa Fe (2005)
43. Simpson, T.K.: Maxwell's Mathematical Rhetoric: Rethinking the Treatise on Electricity and Magnetism. Green Lion Press, Santa Fe (2010)
44. Smith, C., Wise, M.N.: Energy & empire: A biographical study of Lord Kelvin. Cambridge University Press, Cambridge (1989)
45. Thomson, W. (Lord Kelvin): On the uniform motion of heat in homogeneous solid bodies, and its connexion with the mathematical theory of electricity. In: Thomson, W. (ed.) Reprint of Papers on Electrostatics and Magnetism, pp. 1–14. Macmillan & Co, London. (1842/1872) (Originally published 1842)
46. Thomson, W. (Lord Kelvin): On the mathematical theory of electricity in equilibrium I. On the elementary laws of statical electricity. In: Thomson, W. (ed.) Reprint of Papers on Electrostatics and Magnetism, pp. 15–37. Macmillan & Co, London. (1845/1872) (Originally published 1845)

47. Thomson, W. (Lord Kelvin): A mathematical theory of magnetism. In: Thomson, W. (ed.) Reprint of Papers on Electrostatics and Magnetism, pp. 340–425. Macmillan & Co., London (1849/1872) (Originally published 1849)
48. Turner, M.: Cognitive Dimensions of Social Science. Oxford University Press, Oxford (2001)
49. Tweney, R.D.: Inventing the field: Michael Faraday and the creative "engineering" of electromagnetic field theory. In: Weber, R.J., Perkins, D.N. (eds.) Inventive Minds: Creativity in Technology, pp. 31–47. Oxford University Press, Oxford (1992)
50. Tweney, R.D.: Representing the electromagnetic field: how Maxwell's mathematics empowered Faraday's field theory. Sci. Educ. **20**(7–8), 687–700 (2011)
51. 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, pp. 406–435. Oxford University Press, Oxford (2012)
52. Warwick, A.: Masters of Theory: Cambridge and the Rise of Mathematical Physics. University of Chicago Press, Chicago (2003)
53. Weiskopf, D.A.: Embodied cognition and linguistic comprehension. Stud. Hist. Philos. Sci. **41**, 294–304 (2010)

Bohr's Theory of the Hydrogen Atom: A Selective Realist Interpretation

Michel Ghins

Abstract This paper proposes a reconstruction of 1913 Bohr's theory of the hydrogen atom in the framework of the model-theoretic approach of theories. On the basis of this reconstruction, it is argued that Bohr's theory is not internally inconsistent and can't be qualified as fictitious. Then, a selective realist interpretation of Bohr's theory can be defended according to which electrons occupy energy levels. An agnostic attitude however is recommended as far as the electron's trajectories are concerned.

Bohr's theory of the hydrogen atom has been the subject of extensive discussion in philosophy and history of science. Yet, no detailed reconstruction of Bohr's theory and way of proceeding in the framework of the model-theoretic approach of theories has been provided so far.[1] Here, I propose such a reconstruction, but I also give its due to the role of propositions as important ingredient of scientific theories. In fact, theories consist in mathematical structures, namely models, and also in propositions made true or satisfied by those models. Furthermore, a scientific theory provides useful information about entities in the world only if the targeted entities do possess the properties referred to in the theory.

On the basis of my proposed reconstruction, I will attempt to substantiate the following two claims. First, contrary to common opinion, Bohr's theory is *not* internally inconsistent and cannot then readily dubbed a fiction. Second, a

I wish to thank Alisa Bokulich, Jean Bricmont, Harvey R. Brown, Steven French and Peter Vickers for their comments on a previous draft of this paper.

[1] But see my recent paper [1].

M. Ghins (✉)
Centre Philosophie des Sciences et Sociétés (CEFISES), Institut Supérieur de Philosophie,
Place du cardinal Mercier, 14 B-1348 Louvain-la-Neuve, Belgium
e-mail: michel.ghins@uclouvain.be

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_22, © Springer-Verlag Berlin Heidelberg 2014

selective realist interpretation of Bohr's theory, which remains non-committal about the existence of electron's trajectories, can still be defended.

1 Models and Theories in Physics

In the following discussion, I will rely on a *synthetic* view of theories according to which a theory is made up of models and propositions, such as laws, that are satisfied or made true by those models. If what I call *ontic propositions* or judgements [2], that is, propositions[2] that affirm that some entities in the world possess the properties involved in the theory, are true, then the theory is applicable to those entities. And the theory can be said to be (approximately) true of those entities if their properties are related in the way described by the theory. Let us briefly explicate these notions.

According to Tarski [3] and Suppes [4], a model is, in the first place, a mathematical *structure*, that is, a set of individuals on which some relations hold. A structure M thus consists in two ingredients, namely a *domain* D of individuals and a set or relations r_1, r_2,… r_n on these individuals:

$$M = <\mathrm{D},\ \mathrm{r}_1,\ \mathrm{r}_2,\ \mathrm{r}_3\ldots>\ \text{ or } M = <\mathrm{D}, \mathrm{r}_\mathrm{i}>\ (1 \leq \mathrm{i} \leq \mathrm{n})$$

For a structure to become a mathematical model it must satisfy or make true some propositions. Mathematically, a model always is a model *of* a set of propositions. The propositions and the models that satisfy them constitute a *theory*.

In physics, models are used to *represent*. If representation is to be successful, the user of the model must establish a specific kind of representational relation between the representing structure and what it represents. A *sine qua non* condition for the success of mathematical representation in physics (and, in general, in the empirical sciences) is to look at the entities (things, processes or whatever) in the world as *systems*, that is, as domains of individual *properties* structured by specific relations. To take a familiar example, the systems of cartesian mechanics are the structures composed of the properties of "figure and motion" organized by geometrical relations.

Success in scientific representation presupposes an initial abstracting attitude in which the scientist looks at the entities in the world as *structures of properties*. This I called the *primary* or *original* abstraction [5]. The following abstracting step, or *secondary* abstraction, consists in paying attention only to the properties that are judged to be relevant according to some interests (such as positions and velocities) and their numerical values.

Together with Patrick Suppes [4], Newton da Costa and Steven French [6], among others, I contend that the construction of a *homomorphism* is a *necessary*

[2] I make the customary distinction between statements and propositions: propositions are the semantic content of statements.

(*albeit* certainly *not* sufficient) condition for successful representation in science [2, 7]. A homomorphism is a function between two domains that leaves invariant the relations holding on these domains. In this sense, a homomorphism preserves the form. Structures that are homomorphic are said to be *structurally similar*. An *isomorphism* is a one–one homomorphism.[3] A mathematical structure acquires the status of model of a real (or possible) system only if we have constructed a homomorphism or an isomorphism between the model and the (possible) targeted system. Such a requirement is an extremely weak constraint on representation. We can construct a homomorphism between a mathematical structure and any entity in the world. We only have to select some properties of the entity and some relations among those properties. Then, we construct a homomorphism between the mathematical structure and the structure of the entity's properties. In science we of course strive to construct informative representations. To achieve this goal some more restrictive conditions will be required.

According to a selective scientific realist, some propositions of the theory are true, or at least approximately so, and some entities in the world possess the properties referred to in the theory. Entity realism and proposition realism go hand in hand. A realist interpretation of a theory rests on the existence of properties, that is, on the truth of ontic statements which assert that entities which possess those properties exist. Thus to assert that electrons exist is tantamount to saying that there are entities in the world which have a charge, a spin, a mass, an energy and may be properties of velocity and position at all times that make up what is called a "trajectory". Furthermore, those properties—more accurately, the values for those properties (which also are properties…)—are organized by means of relations that constitute a model. Such a model correctly represents a real system of properties if it is true that some entities in the world possess those properties and that the propositions describing the relations between those properties are true. For example, if it is true that there are entities in the world, such as planets, that are located at some distance to a centre and have an orbital frequency and if some mathematical relation holds between the values of orbital frequencies and distances, then a model in which these quantitative properties are structured by means of this relation correctly represents a real system.

2 Modelling an Observable Entity: The Sky

In order to model the observable entity we call the sky, we must first look at the sky as a system and not as a celestial vault of stupendous beauty. This is the primary or original abstracting move. We call some moving bright spots "planets"

[3] Two structures $M = <\mathrm{D}, \mathrm{r}_\mathrm{i}>$ and $M^* = <\mathrm{D}^*, \mathrm{r}^*_\mathrm{i}>$ are isomorphic if and only if there exists a one–one function *f* such that for all r_i and for all n-uple $(\mathrm{a}_1, \ldots \mathrm{a}_\mathrm{n})$ of elements of D which stand in relation r_i, there exists a n-uple $(\mathrm{a}^*_1, \ldots \mathrm{a}^*_\mathrm{n})$ of elements of D* which stand in a r^*_i such that $\mathrm{a}^*_1 = f(\mathrm{a}_1), \ldots, \mathrm{a}^*_\mathrm{n} = f(\mathrm{a}_\mathrm{n})$ ([4], pp. 54–57).

and other bright spots that don't seem to move one with respect to another "fixed stars". In the secondary abstracting move, we pay attention to the changing spatial relations of the five visible planets with respect to the fixed stars. More specifically we are interested in the distances of the planets from the sun and also in their orbital frequencies, that is the number of their complete revolutions in some unit of time. We see thus the sky as a *system* the elements of which are the *observed* distances to the sun and orbital frequencies of the five planets. Let me emphasize that the elements of the system considered are *not* the planets, but some of their *properties*, namely their distances to the sun and their orbital frequencies, while other properties such as their brightness and colour are left aside.

The distances to the sun and orbital frequencies of the five visible planets are observable with the naked eye. These properties can be approximately ordered, and thus structured, according to their sizes observed from the earth. (In remote antiquity, astronomers have already observed that the orbital frequency of a planet is inversely proportional to its distance to the sun). By doing so, we have constructed an *observable* or *phenomenal* structure *O*.

Aiming at precision, we use measuring apparatuses to determine the distances *a'* and the orbital frequencies *ω'* of the planets (I'll use a " $'$ " to refer to data, *i.e. measured* values, as opposed to theoretically *calculated* values). Now, we are confronted with the perennial astronomical problem of finding a mathematical function which fits the measurement outcomes, *i.e.* the *data*. Assuming that planets move on circles[4] and placing a hypothetical observer at the centre (the sun) of those circles, we can construct a data structure *K* the domain of which consists in the distances and frequencies measured at the centre. (These data can be obtained from measurements performed on earth by means of a simple mathematical formula). The measured distances and frequencies approximately verify the following equation:

$$\omega' = 2\pi/\left(a'^{3}k\right)^{1/2} \tag{1}$$

This is Kepler's third law (*k* is approximately a constant). The *data structure K* consists in the measured distances and frequencies organized by Eq. (1). Then, we establish a homomorphism between the phenomenal structure *O* and the data structure *K* and claim that the latter represents the latter. In other words, the data structure *K* is a data *model* that represents the phenomenal system or structure *O*.

A model is useful if it conveys interesting information about some entities in the world. Statements such as "The orbital frequency of Mercury ω'_1 is equal to $2\pi/(a'^3_1k)^{1/2}$" are (approximately) satisfied by this data model.[5] To get information

[4] Remember that we have Bohr's model (which assumes that electrons move on circular orbits) in mind. Later on, Bohr's model was refined by Sommerfeld who introduced elliptical orbits.

[5] This appears to be redundant, even trivial. But see Suppes' very simple example of a domain D of two natural numbers 1, 2 organized by the order relation $\geq$, namely $D = <\{1, 2\}, \geq>$ which satisfies propositions such as $2 \geq 1$ ([4], p. 26).

on the planetary motions we must know the meanings of the numerical symbols: they refer to the orbital frequencies and distances of the planets. More generally, we must know the *code* used in representing. Besides, the measurement data must be sufficiently *accurate*. The data model conveys information about what is going on in the world only if some propositions are true, otherwise the data model represents only a simply *possible* phenomenal system. If predicative propositions such as "The orbital frequency of Mercury is ω'_1" and "The orbital frequency of Mercury is a'_1" are true, the data model correctly represents the phenomenal system. Such predicative propositions are *not* made true by the data model but they provide the grounds for its accuracy. Their truth warrants that the data model possesses some informative content. I call predicative propositions of this type *ontic propositions* since they attribute properties to some concrete real entities, such as planets. If these ontic propositions are true, then it is also true that the real properties denoted by ω'_1 and a'_1 stand in relation (1).

So far, we have constructed structures that are merely kinematical. Isaac Newton introduced dynamical quantities, such as mass and force, in order to provide an explanation of Kepler's laws. The gravitational force is a central force proportional to the masses involved, which decreases with the square of the distance between them. Using Newton's laws, supplemented with initial conditions, we are in a position to calculate the properties, namely the radiuses a and orbital frequencies ω, for all possible stationary orbits of a two-mass system. To each orbit is associated a specific energy level W. We then construct a *theoretical model* H the domain of which contains the *calculated* energy levels of the five planets, their masses m, their orbital frequencies ω and distances to the centre a, which (exactly) verify the following equation:

$$W = 2\pi^2 m\omega^2 a^2 \tag{2}$$

In the next step, we construct a homomorphism from H into K that maps the theoretical orbital frequencies and distances in H into approximately equal measured orbital frequencies and distances in K. The theoretical model H can be merged in a larger structure N, which contains all possible trajectories and energy levels for the two-mass system and also Newton's laws. In van Fraassen's terms, H is an *empirical substructure* of N, but it is also a *theoretical* structure since it is included in a larger theoretical structure. The wider model N is said to be *empirically adequate*. In this way, the phenomenal structure O has been successfully *embedded* in the wider Newtonian model N.[6]

[6] I added the notion of phenomenal or observational model as the intermediary link between phenomena and data models to van Fraassen's model-theoretical approach, which, as the reader noticed, is one of my main sources of inspiration (see [7]).

My construction can be summarized in the following way:
Sky (phenomenon)
↓ abstraction
Phenomenal structure *O* of *observed* planetary orbital frequencies and distances
$\sim$ isomorphism
Data model *K* of *measured* frequencies and distances
$\sim$ homomorphism
Empirical structure *H* of *calculated* frequencies and distances
$\cap$ set-theoretical inclusion
Newtonian theoretical model *N*

Here, the theoretical Newtonian model corresponds to a two-mass system in which one mass is much bigger than the other. The domain of this Newtonian model contains all the continuous possible energy levels of the smaller mass. The *differences* between two possible energy levels are obtained by simple calculation. This will be relevant when we get to the Bohr model.

3 Modelling the Unobservable: The Hydrogen Atom

Modelling *un*observable entities is of paramount importance in physics. On this, there is little disagreement between realists and antirealists. Scientists typically use well-established theories as heuristic guides for constructing novel models. As Mary Hesse [8] emphasized, this demarche assumes some kind of resemblance, or *positive analogy*, between the newly investigated entity and an already successfully modelled entity. Specifically, some properties and relations must be assumed to be shared by both entities. In attempting to model the hydrogen atom, Bohr [9] had good reasons to believe, on the basis of the scattering experiments with alpha particles performed by Geiger and Marsden, that the atom and the sky are analogous in some relevant respects. In accordance to what Perrin envisioned as soon as in 1901 and Rutherford later hypothesized, just as the sky can be partially seen as a system of planets revolving around the very massive sun, the atom can be considered as a system of electrons revolving around a very massive nucleus. Some properties are possessed by both the planets and the electrons, namely their distances to the centre and orbital frequencies. Moreover, since electrons and nucleus have opposite (equal) charges, they are bound to the nucleus by the Coulomb static force that has the same mathematical form as the gravitational force (in atoms, the gravitational force is much weaker than the electromagnetic force and can be neglected). Thus, the truth of ontic propositions attributing some relevant properties of the same kind to planets and electrons is assumed.

Treading in Bohr's footsteps, I model the hydrogen atom by constructing the structure of the properties of all possible stable circular orbits for a single electron revolving around the nucleus. These properties are the distances to the nucleus, the orbital frequencies and the energy levels. This purely theoretical construction deals with quantities that are *calculated* by means of classical laws. Then, it is easy to construct a homomorphism from a planetary structure H_K into an atomic structure H_R[7] which maps energies, distances and frequencies into energies, distances and frequencies and which leaves Eq. (2) invariant. Let me stress that H_K and H_R are not only structurally similar but they are also alike. They resemble each other in the sense of Mary Hesse's positive analogy.

Bohr now encounters a serious obstacle. Maxwell's laws of classical electrodynamics stipulate that an accelerated orbiting electron emits electromagnetic radiation. As a consequence, a revolving electron should continuously fall towards the nucleus while emitting a radiation the frequency of which is also continuous. On the contrary, the electromagnetic emission and absorption spectra provide ample evidence that stable atoms do not emit radiation and that the emission and absorption spectra of excited atoms are not continuous but discrete. Experimental data manifest a blatant contradiction between Rutherford's conception of the atom and classical electromagnetism. Taking a decisive clue from Balmer's formula,[8] Bohr boldly assumes first that an electron can revolve on closed stable orbits without emitting radiation and second that an electron emits or absorbs radiation only when jumping from one stable orbit to another. Moreover, since the atomic spectra exhibit sharp discontinuous lines, Bohr supposes that only some stable orbits are permitted. In the case of the hydrogen atom, Bohr assumes that its single electron moves on specific allowed *circular* orbits possessing discrete kinematical properties, namely their *radiuses* a_n, *orbital frequencies* ω_n and *energy levels* W_n (the allowed orbits are indexed by the natural number n).

Being aware of the importance of Planck's constant for the behaviour of atomic systems and guided by the necessity of recovering Balmer's formula, Bohr introduces the following non-classical condition for the quantization of energy as a function of the frequency ω_n Planck's constant h and the quantum positive natural number n:

$$W_n = nh\frac{\omega_n}{2} \tag{3}$$

This quantization condition is absolutely central to Bohr's reasoning. Yet, this is his *sole* non-classical assumption. At this point, we have what Mary Hesse calls a *negative analogy*. Whereas planets can occupy a continuous array of stationary orbits, electrons are allowed to move on some selected orbits only. After stating the quantification condition, Bohr's reasoning is purely mathematical without resorting to other non-classical suppositions. An energy level W_n is a function of the mass m of the electron, its charge e, Planck's constant h and the quantum number n:

$$W_n = \frac{2\pi^2 me^4}{n^2h^2} \tag{4}$$

[7] The subscripts K and R are chosen in honour of Kepler and Rutherford.

[8] Bohr is quoted as having repeatedly said: "As soon as I saw Balmer's formula, the whole thing was immediately clear to me." (Rosenfeld's Introduction to reprinting of Bohr's papers [10], p. xxxix).

The kinematical quantities, that is, the orbital frequency ω_n and the radius a_n, have been eliminated: they don't appear in (4). We are now in a position to construct a theoretical model H_B the domain of which contains energies and natural numbers structured by Eq. (4). Then, we establish a homomorphic function f_B from H_B into H_R such that the couples (W_n, n) satisfying Eq. (4) are sent to triples (W, a, ω) satisfying Eq. (2) in such a way that the ordering between the energy levels is preserved. My reconstruction of Bohr's demarche can be summarized thus:

Two-mass system		Hydrogen atom
↓ abstraction		↓ abstraction
Keplerian model		Rutherford's model
H_K = <$\mathrm{H_k}$, $\mathrm{h_k}$>	∼	H_R = <$\mathrm{H_R}$, $\mathrm{h_R}$>
	Homomorphism (positive analogy)	
		∼ f_B
		H_B = <$\mathrm{H_B}$, $\mathrm{h_B}$>

In order to gain experimental access to the electron's energy levels, Bohr assumes that when an electron jumps from orbit n' to orbit n closer to the nucleus (with n smaller than n'), a single monochromatic quantum of radiation hv is emitted. The energy of this quantum of radiation is equal to the *difference* between the energies associated with these orbits:

$$W_n - W_{n'} = \frac{2\pi^2 me^4}{h^2}\left(\frac{1}{n^2} - \frac{1}{n'^2}\right) = hv \tag{5}$$

and

$$v = \frac{2\pi^2 me^4}{h^3}\left(\frac{1}{n^2} - \frac{1}{n'^2}\right) \tag{6}$$

Balmer's formula is a particular case of (6) for $n = 2$. The values of the emitted frequencies are not directly observable, but colours are and can be analysed with the aid of a spectrometer. From the series of lines on a photographic plate, we obtain a *data model* D_v of frequencies and natural numbers structured by formula (6) with the condition $n < n'$. Such a data model is homomorphic to a model of *differences* of energies which can immediately be constructed from H_B. For n' infinite, we can construct a data model $D*_v$ homomorphic to the model H_B. For an electron falling from infinity, that is, the binding of a free electron, the energy levels in H_B are mapped into electromagnetic frequencies such that pairs (W_n, n) satisfying (4) are mapped into corresponding pairs (v, n) verifying (6). If the measured frequencies are approximately equal to the calculated frequencies, we are entitled to claim that Bohr's theoretical model is empirically adequate.

The reconstruction of Bohr's demarche offered above, is based on his epoch-making 1913 papers. (For further discussion, see Ghins [1]).

4 Some Philosophical Lessons

Two interesting philosophical lessons can be drawn from the reconstruction just proposed. First, Bohr's model is not *internally* inconsistent; thus, there is no need to resort to quasi-structures and partial homomorphisms as Newton da Costa and Steven French [6] propose. Second, given its internal consistency, Bohr's model remains open to a realistic interpretation and cannot be immediately interpreted as fictitious. This leaves open a selective realist interpretation of Bohr's model according to which energy levels are real whereas agnosticism with respect to the existence of electron's trajectories is recommended. Let us examine these two issues in turn.

4.1 Inconsistency

Most historians and philosophers contend that Bohr's model of the hydrogen atom is inconsistent. In a sense, they are quite right since Bohr's assumptions conflict with Maxwell's electromagnetism according to which accelerated electrons cannot occupy stable energy levels. Bohr's assumptions also conflict with classical mechanics since the latter allows a continuum of energy levels, in contradiction with the quantization condition (3). Several logicians and philosophers [6, 11, 12] even praise inconsistency as virtuous, since the empirical success of Bohr's theory allegedly proves that paraconsistent logic is useful in some instances of scientific modelling. However, together with Peter Vickers[9] and others, I will briefly argue that Bohr's model is *not internally* inconsistent.[10]

At the end of the day, in a physical theory, only the models and the equations satisfied by those models are relevant. The equations used by Bohr to deduce the electron's energy levels and Balmer's formula have not been proved to be inconsistent: it would behove to those who might wish to advocate their inconsistency to produce a proof of their contention. Granted, Bohr's crucial assumption (3) contradicts classical physics, but as Vickers and others (Bartelborth [14, 15]; Hendry [16]; Hettema [17] aptly point out, *external* inconsistency must be sharply distinguished from *internal* consistency. If we take Bohr's mathematical theory of the hydrogen atom in itself, leaving aside its acknowledged conflict with classical physics, the logical consistency of the relevant equations can be safely assumed. The main charge of inconsistency is based, as we saw, on Bohr's assumption that the electron does *not* emit electromagnetic radiation while orbiting on an a stationary orbit n (especially on the lowest orbit for which n equals 1, *i.e.* the ground state).

[9] For a detailed discussion of the inconsistency charges that have been levelled against Bohr's model, see Vickers [13], Chap. 3).

[10] For more discussion on this, see (Ghins [1]).

> (…) the assertion that the ground state was stable, so that an electron in such a state would not radiate energy (…). *This* is the central inconsistency. (Da Costa and French [6], p. 91)

What Bohr does is to separate Coulomb's law for electrostatics from the rest of classical electromagnetism. Having done this, he proceeds with his deductions (see Hendry [16] and Norton [18]). Then the contradiction with other laws of classical electromagnetism, *albeit* certainly problematic, becomes external to Bohr's theory of the hydrogen atom. Thus, given the internal consistency of Bohr's theory, there is no need to resort to quasi-structures as Da Costa and French propose ([6], p. 91). In a quasi-structure, three kinds of relations are distinguished. R_1 are the relations that we assume to hold between the elements of the domain; R_2 are the relations that we suppose not to hold between these elements; R_3 are the relations about which we are non-committal with respect to their holding or not between the elements of the domain (Da Costa and French [6], p. 19). It can readily be shown that propositions which contradict each other can be satisfied by an adequately constructed quasi-structure.[11] However, in my reconstruction of Bohr's theory, some parts of electrodynamics are simply ignored since they play no role in the deduction of the energy levels. This is a consequence of the abstraction procedure in play in any modelling activity. The mathematics of Bohr's theory just remains non-committal about the truth of some laws of electrodynamics when dealing with bound electrons. Obviously, Bohr doesn't need to take into account the classical law of electromagnetic emission by accelerated charges, since he postulates that the electron doesn't emit radiation when on a stationary orbit. Robert Millikan's contemporary reaction to Bohr's theory is worth quoting here:

> Bohr's first assumption (…) when mathematically stated takes the form $e^2/R^2 = (2\pi\omega)^2mR$ in which e is the charge of the electron, ω the orbital frequency, and m the mass of the electron. This is merely the assumption that the electron rotates in a circular orbit… The radical element in it is that it permits the negative electron to maintain this orbit or to persist in this so-called "stationary state" without radiating energy even though this appears to conflict with ordinary electromagnetic theory. (Millikan 1917, quoted by vickers [19], p. 247)

Such reaction shows that some physicists were perfectly aware that Bohr constructed his theory by putting to work only some parts of classical theory while not taking some other parts into account. This doesn't imply that the external contradiction with Maxwell's electromagnetism was ignored. On the contrary, such contradiction was deemed to be a serious defect of the theory, including by Bohr himself who judged it "provisional" (Pais [20], p. 155). In a December 1913 talk, Bohr said:

> You understand, of course, that I am by no means trying to give what might ordinarily be described as an explanation… I hope that I have expressed myself sufficiently clearly so that you have appreciated the extent to which these considerations conflict with the admirably coherent group of conceptions which have been rightly termed the classical theory of electrodynamics. (Pais [20], p. 155)

[11] Some telling examples are discussed by Vickers [19].

Now as then, physicists strive, and rightly so, to construct unified theoretical ensembles that are immune of contradiction. When a successful theory clashes with well-established background laws, scientists are confronted with a *conceptual problem* [21]. To retrieve consistency, scientists devote considerable efforts in revising parts of accepted physics. The history of science provides ample evidence for the fecundity of such efforts. In the present case, these efforts gave birth to standard quantum mechanics. Praising the virtues of inconsistency in some instances (if there are any) should not impair the pursuit of consistent theories, a pursuit which has been highly beneficial to scientific progress.

4.2 A Selective Realist Interpretation

In what follows, I will try to answer the following question: to what extent are we entitled to believe in a selective realist interpretation of Bohr's theory of the hydrogen atom? This is an epistemological issue. Thus, I will be concerned with the arguments in favour of the existence of some specific properties, the truth of some propositions and the accuracy of some representations or models.

As we saw, a scientific theory is a class of models together with the set of propositions that are satisfied by them. Taken in itself, a scientific theory is independent from the existence of any entity in the world. Classical mechanics can be exposed in textbooks irrespective of the actual existence of mechanical systems, just as Euclidean geometry can be developed independently of the existence of rigid rods. Very plausibly, Frederick Suppe [22] has defended a counterfactual account of theories according to which scientific laws are not committed to the existence of entities but solely assert that if entities of a specified kind existed, then their behaviour would be correctly described by some specific mathematical equations. For sure, a theory must at least leave open the possible existence of some systems in the world. For this to be the case, the theory must be logically consistent. Consistency has to do with the truth of propositions. Propositions that negate each other cannot both be true. An electron cannot emit and also not emit radiation at the same time. But, since there is no cogent reason to question the internal consistency of Bohr's theory, a serious obstacle in the way of its selective realist interpretation is removed. If Bohr's theory of the hydrogen atom was inconsistent, then it could be judged to be a fiction right away in the sense that a fiction must at least be something that we have good reasons to believe that it is false (Bokulich [23], p. 137 and [24], p. 92).

In the first place, we may ask if the entities of which Bohr's theory speak, namely electrons, exist. In other words, are there in the world entities which possess specific properties such as numerical values of mass and charge. The achievements of Bohr's theory, no matter how spectacular, do not provide sufficient grounds for believing that electrons exist. Bohr's deduction of the frequencies of spectral lines of the hydrogen atom (and the ionized helium) but also his prediction of the correct value for the Rydberg constant (a fact which strongly impressed Einstein) carry no obligation to commit ourselves to the existence of unobservable entities, such as electrons.[12]

Electrons are entities which are endowed with some properties by our theories only and not by direct observation. To be entitled to believe in the existence of electrons, diverse independent experimental methods of measuring their properties must deliver concordant quantitative results. This criterion is inspired from the manner in which we justify our beliefs in ordinarily observed entities. We feel confident about the existence of a specific observed entity, such as a sample of wine or water, when we are able to perform a diversity of observations from different perspectives that resort to different perceptual modalities (seeing, tasting etc.) and when these observations deliver approximately equal results (see Ghins [26] and [27], p. 108). In 1913 [28], Jean Perrin famously managed to convince the scientific community that atoms exist by showing that thirteen methods of measuring Avogadro's number give approximately equal values.[13] In the same manner, the quantitative results obtained by independent experimental methods provide good grounds to believe that entities having a mass m and a charge e equal to 9.11×10^{-31} kg and 1.6×10^{-19} Coul respectively, do exist. Such evidence in favour of the existence of electrons is *external* to Bohr's model. In other words, the truth of ontic propositions asserting that electrons possess mass, charge, and also energy and that their momentum and position can in some circumstances be measured, is established on the basis of evidence that comes outside Bohr's theory of the hydrogen atom.

We now come the following issue: is it reasonable, given the empirical adequacy of Bohr's theory of the hydrogen atom, to believe that an electron has a trajectory, that is, that it has a momentum (or velocity) and a position at all times and that its spacetime line is continuous? In Bohr's theory, the orbital frequencies ω and radiuses a of the trajectories of electrons are easily *calculable* from the experimental values of the emission and absorption frequencies. But these kinematical properties are not experimentally measurable by a variety of independent experimental methods, contrary to what is the case for planets.

Anjan Chakravartty ([29], Chap. 1) proposes to divide unobservable properties into two distinct categories: those which are *detectable* by means of instruments and those which are not detectable but play an *auxiliary* role in the explanation of phenomena. Detectable properties are those, in the words of Chakravartty, "with which we have managed to forge significant causal contact" ([29], p. 60). Accordingly, the scientific realist should be committed to the existence of detectable properties only and remain agnostic about the existence of auxiliary properties. I contend that even more stringent conditions are required in order to justify the legitimacy of existence beliefs: only properties that are measurable by means of distinct and independent experimental procedures that lead to concordant

[12] Several serious objections have been addressed to the so-called no-miracles argument and the truth-tropic strength of inference to the best explanation [25].

[13] It is debatable however that all thirteen methods are independent from each other.

quantitative values (up to an appropriate degree of approximation) can legitimately be attributed to some entities. In other words, a variety of concordant causal contacts must obtain.

Thus, it is reasonable to be committed to the truth of ontic propositions that assert that "There are entities that have a mass equal to 9.11 × 10^{-31} kg and a charge equal 1.6 × 10^{-19} Coul." as we saw. On the other hand, it is more prudent for the realist to refrain from claiming that electrons move on spatiotemporal trajectories, simply because we have not managed, at least so far, to measure their positions and velocities by several independent ant concordant methods and to experimentally determine the properties of their purported trajectories. The positions and velocities of electrons are not detectable. These are auxiliary properties, that may play some kind of explanatory role [24], but there is no convincing reason for believing that electrons possess them. An agnostic attitude *à la van Fraassen* ([30], p. 72) is recommended with respect to these properties. Such austere selective realism does *not* exclude the possibility that further evidence might be provided in the future in favour of electron trajectories.

One might object that the Heisenberg uncertainty inequalities prohibit the simultaneous measurement of positions and momenta. Those inequalities certainly are a consequence of the new quantum mechanics that was developed after 1913 and became standard later on. However, there is no totally convincing argument for interpreting Heisenberg's relations in a strong realist way and to conclude that electron trajectories do not exist.[14] Such a strong position, akin to atheism in theology, is not warranted. According to Bohm's theory, which is empirically equivalent to standard non-relativistic quantum mechanics, electrons move on trajectories, and can even be at rest. (According to Bohm's theory however, the purported electron trajectories in the hydrogen atom do not coincide with Bohr's orbits and cannot generally be identified with classical trajectories). My aim here is not to take sides in a sometimes (too) heated debate, but to point out that Bohm's theory remains at least a possibility and to recommend an agnostic, thus not atheistic, attitude with respect to the existence of electron's trajectories.

What about the existence of energy levels for bound electrons? According to Bohr's theory, differences in energy levels correspond to spectral lines in the emission spectra. The differences of energy levels are immediately *calculable* from the measured frequencies by means of Eq. (5) above. But the energy differences and the energies themselves are not measured in this way, only the electromagnetic frequencies or wavelengths are. Given the amount of experimental evidence and the variety of methods for ascertaining the emission (and absorption) of electromagnetic radiation by atoms, we can safely believe that atoms do emit and absorb electromagnetic energy and, given the principle of conservation of energy, that atoms occupy—quantified—states of energy. Yet, we need further experiments in order to be able to assert that atomic differences of

[14] I here diverge from Bokulich ([23], p. 137) and the standard interpretation of the Heisenberg principle.

energies correspond to different states of the electron. Bohr's theory alone does not adduce sufficient grounds for believing that the nucleus doesn't play a role in accounting for the observed spectra. Such experimental evidence must come from outside. And it does. Collision experiments show that electrons can be expelled and captured again by atoms, and that the various energy states of atoms correspond to essentially different states of energy for electrons.

Summarizing, given its internal consistency, Bohr's theory cannot be judged to be a fiction, that is, something about which we have good reasons to believe that it does not correspond to some real entities in the world. On the contrary, there are experimental grounds in favour of the existence of electrons, namely entities that possess the real properties of mass, charge and energy. Given this, we also have reasons to believe that the mathematical relations verified by the quantitative values of those properties, such as Eq. (4), are true and that the structure H_B can be considered to be a reliable model and partially and approximately represents existing systems in the world.

References

1. Ghins, M.: Bohr's modelling of the atom. A reconstruction and assessment. Logique et Analyse **218**, 329–350 (2012)
2. Ghins, M. Scientific representation and realism. Principia **15**(3), 461–474 (2011) http://www.cfh.ufsc.br/~principi/15-3.html
3. Tarski, A.: Undecidable Theories. North Holland, Amsterdam (1953)
4. Suppes, P.: Representation and Invariance of Scientific Structures. CLSI, Stanford (2002)
5. Ghins, M.: Realism. Entry of the online Interdisciplinary Encyclopaedia of Religion and Science. http://www.inters.org (2009)
6. Da Costa, N., French, S.: Science and Partial Truth. A Unitary Approach to Models and Scientific Reasoning. Oxford University Press, Oxford (2003)
7. Ghins, M.: Bas van Fraassen on scientific representation. Analysis **70**, 524–536 (2010)
8. Hesse, M.: Models and Analogies in Science. University of Notre Dame Press, Notre Dame (1966)
9. Bohr, N. On the constitution of atoms and molecules. Philosophical Magazine 26(6), 1–25; 476–502; 857–875 (1913) (Re-imprinted with an introduction by L. Rosenfeld (1963), Copenhagen: Munksgaard)
10. Jammer, M.: The Conceptual Development of Quantum Mechanics. McGraw-Hill Book Company, New York (1966)
11. Bueno, O.: Why inconsistency is not hell. Making room for inconsistency in science. In: Olsson, E. (ed.) Knowledge and Inquiry: Essays on the Pragmatism of Isaac Levi, 70–86. Cambridge University Press, Cambridge (2006)
12. Priest, G.: Inconsistency and the empirical sciences. In: Meheus, J. (ed.) Inconsistency in Science, 119–128. Kluwer Academic Publishers, Dordrecht (2002)
13. Vickers, P.: Understanding Inconsistent Science. A Philosophical and Metaphilosophical Study. Oxford University Press, Oxford (2013)
14. Bartelborth, T.: Is Bohr's model of the atom inconsistent? In: Weingartner, P., Schurz, G. (eds.) Proceedings of the 13th International Wittgenstein Symposium, HPT (1989)
15. Bartelborth, T.: Kann es Rational Sein, eine Inkonsistente Theorie zu Akzeptieren? Philosophia Naturalis **26**, 91–120 (1989)

16. Hendry, R.F.: Realism, history and the quantum theory: philosophical and historical arguments for realism as a methodological principle. LSE, unpublished PhD thesis (2003)
17. Hettema, H.: Bohr's theory of the atom 1913–1923: a case study in the progress of scientific research programmes. Stud. Hist. Philos. Mod. Phys. **26**, 307–323 (1995)
18. Norton, J.: How we know about electrons. In: Nola, R., Sankey, H. (eds.) Issues in Theories of Scientific Method, 67–97. Kluwer Academic Publishers, Dordrecht (2000)
19. Vickers, P.: Can partial structures accommodate inconsistent science? Principia **13**, 233–250 (2009)
20. Pais, A.: Niels Bohr's Times, in Physics, Philosophy and Polity. Clarendon Press, Oxford (1991)
21. Laudan, L.: Progress and its Problems. University of California Press, Berkeley (1977)
22. Suppe, F.: The Structure of Scientific Theories. University of Illinois, Chicago (1974)
23. Bokulich, A.: Reexamining the Quantum-Classical Relation. Beyond Reductionism and Pluralism. Cambridge University Press, Cambridge (2008)
24. Bokulich, A. Explanatory fictions. In: Suarez M. (ed.) Fictions in Science: Philosophical Essays on Modeling and Idealization, 91–109. Routledge, London (2009)
25. Ghins, M.: Putnam's no-miracle argument: a critique. In: Clarke, S., Lyons, T. (eds.) Recent Themes in the Philosophy of Science: Scientific Realism and Commonsense, Australasin Studies in the Philosophy of Science, **17**, Kluwer Academic Publishers, Dordrecht, 121–138 (2002)
26. Ghins, M.: Scientific realism and invariance. In: Proceedings of the Third SOFIA Conference on Epistemology. Campinas 30 July–1 Aug 1990. Philosophical Issues, Vol. 2: Rationality in Epistemology, pp. 249–262. Ridgeview, California (1992)
27. Ghins, M.: Can common sense realism be extended to theoretical physics? Log. J. IGPL **13**, 95–111 (2005). (Oxford UP)
28. Perrin, J.: Les atomes. Alcan, Paris (1913)
29. Chakravartty, A.: A Metaphysics for Scientific Realism. Knowing the Unobservable. Cambridge University Press, Cambridge (2007)
30. Van Fraassen, B.: The Scientific Image. Oxford University Press, Oxford (1980)

Identifying Adequate Models in Physico-Mathematics: Descartes' Analysis of the Rainbow

Albrecht Heeffer

Abstract The physico-mathematics that emerged at the beginning of the seventeenth century entailed the quantitative analysis of the physical nature with optics, meteorology and hydrostatics as its main subjects. Rather than considering physico-mathematics as the mathematization of natural philosophy, it can be characterized it as the physicalization of mathematics, in particular the subordinate mixed mathematics. Such transformation of mixed mathematics was a process in which physico-mathematics became liberated from Aristotelian constraints. This new approach to natural philosophy was strongly influenced by Jesuit writings and experimental practices. In this paper we will look at the strategies in which models were selected from the mixed sciences, engineering and technology adequate for an analysis of the specific phenomena under investigation. We will discuss Descartes' analysis of the rainbow in the eight discourse of his *Meteorology* as an example of carefully selected models for physico-mathematical reasoning. We will further demonstrate that these models were readily available from Jesuit education and literature.

1 The New Physico-Mathematics

The physico-mathematics that emerged at the beginning of the seventeenth century entailed the quantitative analysis of the physical nature with optics, meteorology and hydrostatics as its main subjects. Isaac Beeckman after his encounter with

A. Heeffer (✉)
Research Foundation Flanders (FWO Vlaanderen), Ghent University, Gent, Belgium
e-mail: albrecht.heeffer@ugent.be

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_23, © Springer-Verlag Berlin Heidelberg 2014

Descartes in 1618, wrote in his *Journal* that there are not many students of physico-mathematics.[1] While they understood physico-mathematics as a new discipline, the term was previously used within the Aristotelian tradition, denoting related disciplines known as mixed mathematics [25]. Although care should be taken with the use of the term (as argued by Maarten Van Dyck [29]), as it seems that its use in the early seventeenth century already reflects an important semantic shift and change in practices. Mixed mathematics functioned as an intermediate between natural philosophy and pure mathematics, but were considered subordinate to them. Rather than considering physico-mathematics as the mathematization of natural philosophy, John Schuster [25] has characterized it as the physicalization of the subordinate mixed mathematics. Such transformation of mixed mathematics was a process in which physico-mathematics became liberated from Aristotelian constraints.

Peter Dear [5] has shown how this new approach to natural philosophy was strongly influenced by Jesuit writings and experimental practices. Representatives of the tradition, such as Mydorge, Descartes, Mersenne and Cassini were educated at Jesuit colleges while others, such as Fabri, Grimaldi and Scheiner were Jesuits themselves. In this paper we will look at the strategies in which models were selected from the mixed sciences, engineering and technology adequate for an analysis of the specific phenomena under investigation. We will discuss Descartes' analysis of the rainbow in the eight discourse of his *Meteorology* as an example of carefully selected models for physico-mathematical reasoning. We will further demonstrate that these models were readily available from Jesuit education and literature.

2 Material Models for Discovery in Natural Philosophy

Many studies since the late 1990s have demonstrated the importance of models in different aspects of scientific practice. Basic scientific processes such as discovery, explanation, simulation, representation, and experimental design have all been approached as forms of model-based reasoning. These models have been studied from philosophy of science [13, 21] as well as from cognitive science [19, 20]. The models we are interested in are representational models for physical phenomena. Such models typically bear relevant similarities with the phenomena observed, such as shape or elasticity and will respond to certain interactions in ways analogous to the phenomena under observation. Hesse used the term 'material analogies' for such models [16]. However, it is usually through abstraction and

[1] [6], I, 244: "Dicit tamen se nunquam neminem reperisse, praeter me, qui hoc modo, quo ego gaudeo, studendi utatur accuratèque cum Mathematicâ, Physicam jungat. Neque etiam ego, praeter illum, nemini locutus sum hujusmodi studij".

idealization that models reveal their heuristic function for the study of natural phenomena. By abstracting from concrete properties of material models not relevant for a representation of the phenomena, such as friction or deformation, the analogy becomes a formal on in Hesse's terminology. Natural philosophers or scientists arrive at new laws and theories by isolating the relevant features of phenomena in an experimental configuration and excluding the unnecessary and contingent aspects. They do so by carefully selecting those models that facilitate the process of exclusion and isolation. Often a single model will not suffice and the discovery process may therefore involve several related models that allow focusing on one specific aspect more precisely. By using material models, hands-on experience and knowledge embedded as intuitions about our physical interactions with these material objects becomes translated in the target domain to which it is applied.

As a prime historical example for the use of models in scientific discovery we may refer to Descartes' analysis of the optical principles behind the rainbow. While several accounts on the rainbow were available by the 1620s it was Descartes who first succeeded in publishing a satisfactory explanation why the primary and secondary rainbow are always seen under a given angle, why the colors always appear in the same order and why this order is inverted in the second rainbow. He argued his findings by a geometrical analysis of the twofold refraction and single reflection in a raindrop; a corpuscular explanation of light and carefully supported his results by precise quantitative observations. This analysis is presented as the eight discourse of the *Meteorology*, as part of Descartes ground breaking publication *Discours de la méthode* from 1637. Descartes included the discourses on *Meteorology* and *Dioptrics* as an illustration of his general method to arrive at certain and undoubtable knowledge in natural philosophy. As is well established, Descartes' line of explanation in many of these discourses does not faithfully follow his method or his actual path to discovery. (e.g. for the sine law of refraction [14]). Though, for his analysis of the rainbow he explicitly stated that he does. This is one of the reasons for us to take up this example as a case study. Descartes' analysis of the rainbow, conceived as an illustration of his general method, has been the subject of several historical and philosophical studies, notably by Carl Boyer [2], Richard Westfall [27], A. I. Sabra [22], William Shea [24], and what must be the definite account by Jed Buchwald [3]. We will therefore limit this contribution to the specific use of models and the context in which these models became available at the beginning of the seventeenth century. Complementary to several discussions on the use of material and technological models by Descartes (e.g. [12, 28]), we will specifically argue that Descartes could draw his models from Jesuit writings on physico-mathematics as well as popular accounts of disputations organized on these subjects.

3 The Models

Descartes' analysis of the rainbow, which according to his own account follows his path of discovery,[2] is based on a succession of different models to isolate the relevant features of a configuration and exclude the unnecessary and contingent aspects. As the rainbow is a natural phenomenon which can only be observed occasionally, the models facilitate experimental procedures to test and measure a given configuration. We can distinguish the use of five models in the eighth discourse:

3.1 Artificial Rainbows in Fountains

As a natural phenomenon, rainbows can only be observed in specific meteorological circumstances. There has to be a specific combination of rain and sun and the position of the sun should not be higher than 42°. However, the appearance of rainbows can be modeled by artificial rainbows in fountains where they can be observed during a longer period. In the introduction of his treatise, Descartes uses the occurrence of rainbows in fountains as an argument for the reproducibility of this natural phenomenon ([8], VI, p. 325; [9], p. 332):

> First, taking into consideration that this arc can appear not only in the sky but also in the air near us whenever there are drops of water in it that are illuminated by the sun, as experience shows us in certain fountains.

In the last part of the discourse he comes back to fountains and suggests the use of liquids with a higher refractive index for theatrical purposes ([8], VI, p. 343; [9], pp. 344–345).

> And this makes me remember an invention for making signs appear in the sky, which would cause great wonder in those who were ignorant of the causes. I suppose that you already know the method of making a rainbow visible by the use of a fountain. (..) To this it is now necessary to add that there are oils, spirits and other liquids, in which refraction is notably greater or lesser than in common water, and which are no less clear and transparent because of that.

[2] From a letter from Descartes to Vatier, February 1638, [8], I, pp. 559–660: "Nor could I show the use of that method in the three treatises that I included, since it prescribes an order of investigation which is different enough from the order I believed I must use in order to explain them. However, I have given a sample of it in describing the rainbow, and if you take the trouble to reread it, I hope it will satisfy you better that it did the first time". Translation from [10], p. 86.

3.2 The Spherical Glass Modelling a Raindrop

Descartes remarks that a raindrop can be modeled by a spherical flask filled with water ([8], VI, p. 325; [9], p. 332):

> Then, knowing that these drops are round, as has been proven above, and seeing that their being larger or smaller does not change the appearance of the arc, I then took it into my head to make a very large one, the better to examine it. And for this purpose I filled a perfectly round and transparent large flask with water.

Descartes exploits the analogue of a flask with a very large raindrop to empirically come to an estimate on the angle under which the colors appear. He observes that for the color red the angle from the sun to the flask and back to the eye is approximately 42°. Moving the flask upwards makes the color change from red to yellow and then to the other colors of the rainbow. Descartes further determines experimentally that red reappears under an angle of around 52° "but not as brilliant".

3.3 The Prism

In order to achieve a more accurate observation and to eliminate the double appearance of colors (caused by a second internal reflection in the raindrop), Descartes moves to a prism. The prism thus acts as a model in the same way as a spherical flask but allows him to isolate the color separation to a single refraction without reflections. Such a controlled experimental setup allows a more precise observation of the color separation. Descartes wants to know the precise order of colors and to determine what causes the separation of white light into colors. He used what was then known as a triangular glass (*crystallo trigonam*) (MNP in Fig. 1).

3.4 Wooden Bowls

In his corpuscular theory of nature, Descartes conceives of light as the result from movement of small spherical balls (Fig. 2). These corpuscles can be modeled by wooden bowls as known from the game of Kayles (in French: quills) or billiard, or tennis balls as in his first and second discourse on *Dioptrics*, or even grapes as in the first discourse on *Dioptrics*. He uses balls as a model to explain that colors occur by differences in rotational speed of spherical particles. He further exploits tangible experience with these games for hypotheses on the kinetic behavior of corpuscules. A 'top spin' of a wooden ball (as used in billiard or tennis) causes the ball to rotate more than what is needed for a rotational movement on a surface. According to Descartes this corresponds with the color red. A 'slide' or middle hit causes the ball

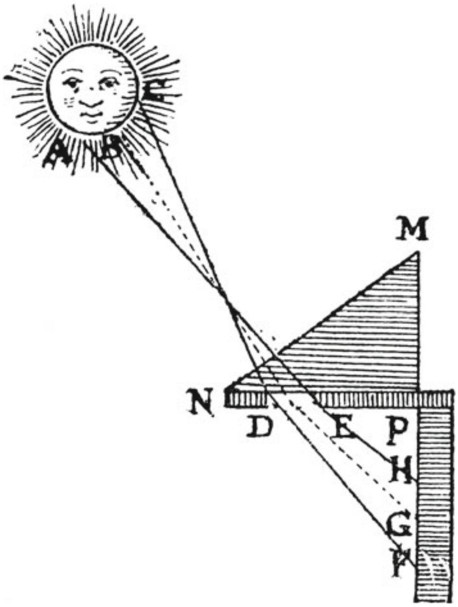


Fig. 1 The prism and aperture from the *Meteorology* by Descartes

to slide and thus rotate less than needed for a rotational movement. This corresponds with blue light. Balls that follow the normal rotational movement on a surface will cause white light. The spin and slide effects on rotating balls are caused by a change of density in optical media and thus occur together with refraction.

3.5 The Aperture

In order to control the spread of colors by the prism, Descartes covers most of the refracting surface of the prism (*DE* in Fig. 1). In doing so he adopts the model of an aperture controlling interactions between neighboring spheres. According to Buchwald [3] who reenacted the historical experiment, Descartes insisted on an aperture to create shade boundaries as he believed that light must interact with dark to generate color. Doing so, Descartes would adopt a transformed version of the Aristotelian theory of color as consisting of a mixture of black and white light. An alternative explanation which we favor is that Descartes was familiar with the *camera obscura* where the size of the aperture influences the quality of the image. It can be empirically established that the ideal width of the aperture is about one hundredth the distance between the aperture and the projected image. It was established only much later, by Lord Rayleigh that the optimal size of the aperture d is given by the formula $d = 1.9\sqrt{f\lambda}$, with f as focal distance and λ as the wavelength of light. For a focal distance of 50 mm this would amount to a very small aperture of 0.32 mm.

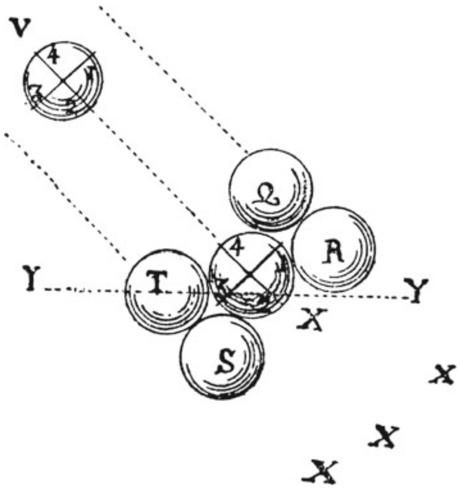


Fig. 2 The corpuscular explanation of color by Descartes

4 Available Models from Jesuit Physico-Mathematics

For each of the models employed, we can identify specific sources that may have been a source of inspiration for Descartes. Jean-Robert Armogathe [1] argued that several of the "observations" by Descartes in the introduction of his eighth discourse on the rainbow stem from the *Meteorologica* published by Froidmont in 1627. While it may be the case that Descartes had read Froidmont's work by 1629, we instead will pursue the remark by Boyer that Descartes was inspired by the anonymous work *Recreation mathematique* (henceforth *RM*), first published at Pont-à-Mousson in 1624.[3] According to Boyer ([2], p. 208) the treatment of rainbows in this book as something which eluded philosophers for a long time posed "a challenge which Descartes could not resist". The book has been wrongly attributed to the Jesuit Jean Leurechon. Elsewhere we have argued that the book was compiled from notes used for the public disputationes at the Jesuit college of Pont-à-Mousson, following closely a compendium of propositions on physico-mathematics authored by Jean Leurechon in 1622, bearing the title *Selectae Propositiones* (henceforth *SP*) [15]. The *RM* and to a lesser extent, the *SP*, were well read, known and discussed by natural philosophers. The *RM* is cited frequently in the correspondence of Beeckman, Descartes, Mydorge and Mersenne. Mydorge's compilation of notes and critical comments was added to the Moreau, 1626 edition and further expanded in the Mydorge edition of 1630. Mersenne, in his *La verite des sciences* of 1625, freely took tens of pages of material from Leurechon's propositions on arithmetic, geometry and music (e.g. pp. 803–812). His only acknowledgement is "comme ont remarqué les mathematiciens du Pont dans leur these de l'an 1622" (p. 803). Mersenne became aware of the *RM* only

[3] For a forthcoming critical edition of the English translation of this book see [15].

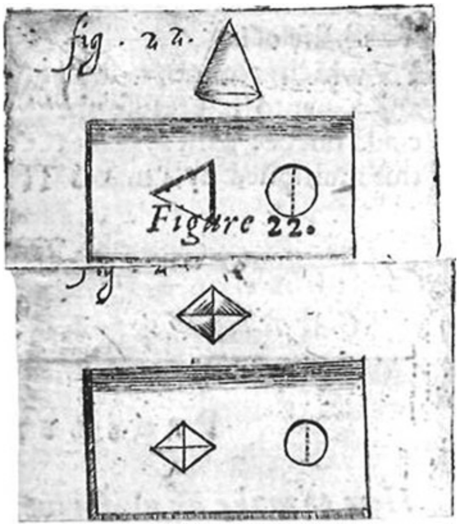


Fig. 3 Conic sections made tangible, from *RM*, problems 22-3.

later. The first reference to the book in his *Correspondence* dates from 27 April 1628. Robert Cornier wrote Mersenne about the volume and that it was expanded with a new third book ([7], II, pp. 83–84). Another early work by Mersenne, *Quaestiones celeberrime in genesis* from 1623 also borrows from the *SP*. The figure on col. 95, is a description of a proposition taken from the section on *Geometry*, prop. X, p. 8.

The *SP* adheres to the Aristotelian view on the mixed sciences. Leurechon describes mechanics as subordinate to geometry and physics, while optics is subordinate to geometry (cited by Dear [5], p. 169. On the other hand, many of the propositions of the book deal with material models, machines, contrivances, mostly borrowed from Cardano's *De subtilitate*, and show a familiarity of engineering treatises from the early seventeenth century. Even subjects in pure mathematics, such as conic sections, are made tangible by two propositions in which inflexible bodies, like cylinders, pyramids and cones, have to be passed through geometrical figures cut in cardboard.[4] These are two of the many propositions reproduced in the *RM* (see Fig. 3).

That the models borrowed by Descartes from *RM* already appear in the witness account of the disputationes held at Pont-à-Mousson should not surprise us. The physicalization of mixed mathematics, as Schusters calls it, was happening already at the Jesuit colleges. Making mathematics tangible was part of a deliberate pedagogical strategy of the Jesuits to allure the nobility to their colleges.

[4] *SP*, *Geometry*, prop. XII, p. 8: "Potest unum & idem corpus solidum ac durum, per tria foramina transmitti, quorum unum sit rotundum, secundum quadratum, tertium ovale, ita ut singula in transitu compleat; potest aliud eandem conditione transmitti per rotundum, ovale & quadr angulare quantaelibet longitudinis: aliud, per rotundum & quandrangulare; alius per rotundum & triangulare duntaxat".

The physicalization of mathematics provided new opportunities for the use of models in physico-mathematics in which Descartes excelled.

In Descartes' account on the rainbow we find that all of the five models he employs also appear in both the *SP* and the *RM*. We list them in the same order.

4.1 Artificial Rainbows in Fountains

The description of the rainbow by Leurechon is rather typical for early seventeenth-century account. We here reproduce the full proposition[5]:

> Refraction and reflection depict various celestial and sublunary appearances, halos, parhelia, paraselenae, rainbows in clouds, in the prism, in glass jars, in multi-faceted crystals, in the arched water of splash fountains, in burning light of a candle. Here one sees with certainty that the colors that appear in these circumstances have a common cause, namely the mixture of the refracted light with the refractive subject; while the different figures that appear are governed simultaneously by the laws of refraction and reflection of the rays of light. It is probable that the moon rainbow can be doubled in the same way as the solar rainbow, yet in such a way that the second bow is not an image produced by a reflection of the first bow.

Problem 46 of the *RM* (44 in the English editions) is an elaboration of this proposition. It includes the "experiment" to fill ones mouth with water and to blow it out through the lips, standing with the back to the sun. The author, in a remark omitted by the English translator, complains about the lack of an explanation on the rainbow ([30], pp. 42–43):

> I am afraid that you will ask me more; about the production, disposition and figure of these colours. I shall reply that they come by reflection and refraction of light, and that is all. (..) And he who said that it is the mirror in which human nature has had a full view of its ignorance has well understood; for all the philosophers and mathematicians who for so many years have been engaged in discovering and explaining its causes, have learned nothing except they know nothing but a appearance of truth

This was the challenge which, according to Boyer [2], Descartes could not let pass by. Descartes also commences his eight discourse with the observation that several phenomena share a common cause. His references to the rainbow in fountains is indeed most likely to be inspired by *RM*. His remark "this makes me

[5] *SP*, *Optics*, prop. XX, p. 35: "Refractio et reflexio, varia coelestibus ac sublunaribus pingit phasmata, halones, parelia, paraselenias, irides in nubibus, in crystallo trigonam, in ampulis vitreus, in polyoptris, in aquam fontium arcuaram et irroratam, in lucernic ardentibus. Ubi certum videtur, colores hic apparentes, causam habere congenerem, puta per mixtionem refractae lucis cum refrigente subiecto; figures autem diversas, incidentium, refractorum simul et reflexorum radiorum legibus gubernari. Probabile est, geminari posse lunarem iridem, sicut et solarem, sic tamen, ut secundaria iris non sit imago alterius ex reflexione producta".

remember an invention for making signs appear in the sky, which would cause great wonder in those who were ignorant of the causes", cited above, is almost a literal quote from the book.[6]

4.2 The Spherical Glass Modelling a Raindrop

The spherical glass is already mentioned by Leurechon (*ampulis vitreus*), as quoted above. In *RM* the glass is described as "a stable and permanent" configuration to observe the colors of the rainbow.[7]

4.3 The Prism

The use of a prism for color separation is described both in *SP* (*crystallo trigonam*) as in *RM* (1633, p. 68): "Or take a triagonal glass, or crystal glass of divers angles; and look through it, or let the beams of the sun pass through it". No further observations are mentioned.

4.4 Colliding Bowls

That mathematical principles govern the behavior of wooden of ivory balls on impact is discussed in *SP*[8]:

> It is a pleasing quality of mathematics to play with balls made of ivory or ebony and the like in such a way as to, [1] by means of the laws of reflection, send to a determinate location one ball striking a plane or two or more planes; or [2] direct one ball, striking the others, in whatever part you will be willing, and even [3] on the condition, if you want, to have assigned lines of various motions, and assigned points of various contacts and collisions, either in a plane or through a circular trajectory.

[6] Compare Descartes, [8], VI, p. 343: "qui pourroient causer grande admiration a ceux qui en ignoreroient les raisons", with one of several similar formulations in *RM*: "avec grand estonnement des assistans particulierement s'ils en ignorent la cause" (1630, p. 73).

[7] From the English edition: 1633, p. 68: "But to have one more stable and permanent in his colours. Take a glass full of water, and expose it to the sun, so that the rays that pass through strike upon a shadowed place, you will have pleasure to see the fine form of a rainbow by this reflection".

[8] *SP*, *Geometry*, XV: "Mathematicae iucunditatis est, in sphaerulis eburneis, aut buxeis, et similibus, ita ludere, ut reflecionem legibus, una in planum, aut duo vel plura plana impingens, ad destinatum locum emittatur: vel ita una sphaerula, caeteras impellens, ad quamcunque volveris partem dirigat, assignatis etiam, si lubet, variorum motuum lineis, variorum contactuum et allisionum punctis, sive in plano, sive per ambitu[m] circuli".

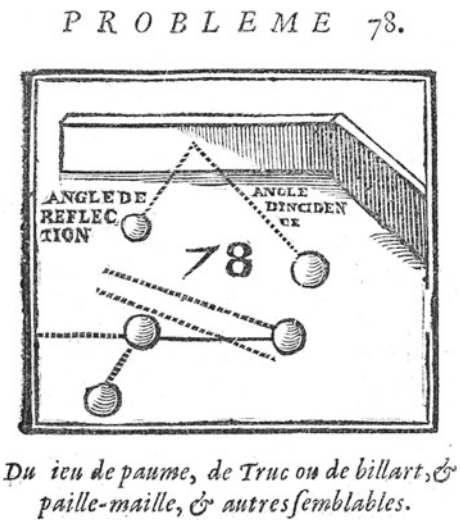

PROBLEME 78.
ANGLE DE REFLECTION
ANGLE DINCIDENCE
78
Du ieu de paume, de Truc ou de billart, & paille-maille, & autres ſemblables.

Fig. 4 A geometrical analysis of billiard from *Récréations Mathématiques*, pr. 78

The planes to which is referred to are the vertical rails, in later versions of the game table replaced by cushions. Leurechon describes three principles which apply to the balls movement without any spin: (1) the law of reflection describing the path when hitting the rail under a given angle, (2) directing the ball to a specific location by hitting the rail more than one time and (3) the effects of colliding balls.

This early reference to ball games should not surprise us. Mac Lean, in a Ph.D. dissertation on the history of the laws of collision [18], describes the popularity of these games in France at the beginning of the seventeenth century. He points at the *RM* as the first work in which ball games are analyzed by the laws of geometry. However, Cusanus used the game of bowling already in his *De ludo globi* of 1463, as a metaphorical tool for expounding his view on the cosmos. Although in this dialogue the bowling ball is most often used as a metaphor, in his argumentation Cusanus explains the laws governing the impetus and movement of the ball. As an example [4], I, p. 23:

> Keep in mind that the motion of the bowling-ball ceases but that the ball remains intact; for the ball has no natural motion but has [only] an accidental and forced motion. Therefore, the ball stops moving when the impetus that was impressed upon it ceases. But if the bowling-ball were perfectly round, then (as was said earlier) its motion would never cease, because a circular movement would be natural to that ball and not at all forced upon it.

The Italian priest and Aristotelian philosopher Antonio Scaino da Salò (1524–1612), published in 1555 a treatise on *giuoco della palla* (*jeu de paume* or tennis), describing the rules of the game and the courts where it is played. He briefly touches upon the relation of the balls movement with the laws of natural philosophy, mainly concerning the principles of ballistics.

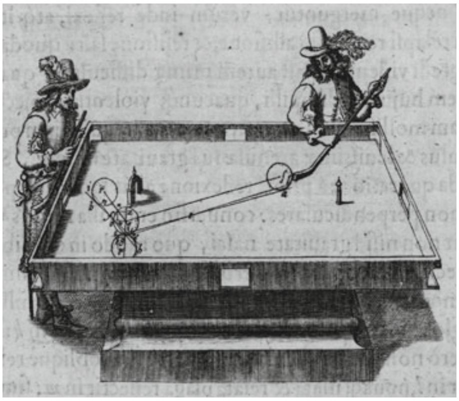


Fig. 5 A geometrical analysis of billard from Johannes Marcus's *De Proportione Motus*, 1636, p. 105

Problem 78 of *RM*, probably contains the first diagram used for a geometrical analysis of a ball game (see Fig. 4). Here we find the three rules described by Leurechon applied to the game of billiard.[9]

> And the first maxim is thus. When a bowl toucheth another bowl or when a trap stick striketh the ball, the moving of the ball is made in a right line, which is drawn from the centre of the bowl by the point of contingency. Secondly, in all kind of such motion, when a ball or bowl rebounds, be it either against wood, a wall, upon a drum, a pavement, or upon a racket, the incident angle is always equal to the angle of reflection.
>
> Now following these maxims, it is easy to conclude:
>
> First, in what part of the wood or wall one may make the bowl or ball go to reflect or rebound, to such a place as one would.
>
> Secondly, how one may call a bowl upon another, in such sort that the first or the second shall go and meet with the third, keeping the reflection or angle of incidence equal.
>
> Thirdly, how one may touch a bowl to send it to what part one pleaseth, such and many other practices maybe done. At the exercises at kayles there must be taken heed that the motion slack or diminish by little and little, and may be noted that the maxim of reflections cannot be exactly observed by local motion, as in the beams of light, and of other qualities, whereof it is necessary to supply it by industry or by strength otherwise one may be frustrated in that respect.

Descartes explains in his discourse on refraction of the *Dioptrique* the causes of refraction by means of an idealized tennis ball (*jeu de paume*) hitting a piece of loosely woven cloth. The second discourse on refraction does not mention a tennis ball explicitly but Descartes refers to the game in the first discourse by describing

[9] From the English edition which does not mention billiard but tennis and trap-ball. This reference to the game of trap-ball predates the earliest entry in the Oxford English Dictionary (1658). The figure is taken from the author's copy of the 1672 Lyon edition. Johannes Marcus expands on these rules in a later treatise of 1636 (see Fig. 5). The earliest depiction of tennis is shown in Fig. 6.

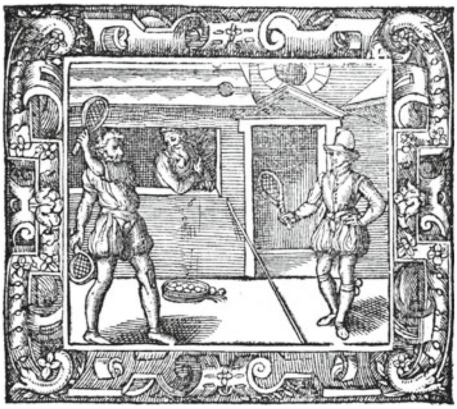


Fig. 6 The earliest depiction of tennis from Sambucus, *Emblemata*, 1564, p. 133

the possible effects on the motion of the ball impacting on a surface. He further explains that the different effects the ball or corpuscule, correspond with the different colors of light ([8], VI, pp. 90–91):

> Those who play tennis can prove this sufficiently, when their ball encounters uneven ground, or else when they hit it obliquely with their racket, which they call, I believe, cutting or grazing (..) Some reflect these rays without causing any other change in their action, such as those which we call white, and the others carry with this reflection a change similar to that which the movement of a ball receives when we graze it, such as those which are red, or yellow, or blue, or any other such color.

In the fifth discourse of the *Dioptrique* on the anatomy of the eye and in the *Meteores* on the rainbow, the acceleration and slip of wooden balls further serves as a model for the behaviour of spherical corpuscules and the causes of color.

The model of the tennis ball further plays a role in the *experimentum crucis* of 1666 by Newton where he describes to the Royal Society his discovery of the cause for chromatics abberation in lenses[10]:

> Then I began to suspect, whether the Rays, after their trajection through the Prisme, did not move in curve lines, and according to their more or less curvity tend to divers parts of the wall. And it increased my suspition, when I remembred that I had often seen a Tennis ball, struck with an oblique Racket, describe such a curve line.

4.5 The Aperture

The the *camera obscura* is described as "that dioptrical instrument" in proposition XIX of the part on *Optics*. The aperture is called a (*pupillae foramen*). Leurechon further describes the use of a concave lens to reinvert the image and draws the parallel with the eye: "for philosophers, it is a fine secret to explaining the organ of

[10] *Philosophical Transactions of the Royal Society*, No. 80 (19 Feb. 1671/2), pp. 3075–3087.

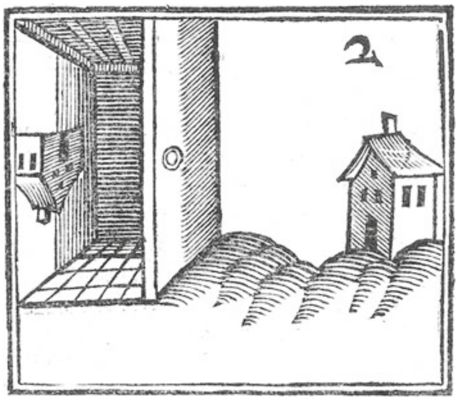


Fig. 7 The camera obscura from *Récréations Mathématiques*, problem 2

the sight, for the hollow of the eye is taken as the close chamber, the ball of the apple of the eye, for the hole of the chamber, the crystalline humour as the small of the glass and the bottom of the eye, for the wall or leaf of paper". The Jesuit must have been familiar with Kepler's *Paralipomena* in which the retinal image theory was first expounded. An elaborated version appears as problem 2 in *RM* (see Fig. 7). Both specifically mention to experiment with the aperture to achieve the best results. The translator of the English edition recommends an aperture of the size of a six pence coin.

Descartes, in the fifth discourse on *Dioptrics* describes his experiences with the *camera obscura* in relation to the eye. Even without the use of a lens "certain images will certainly appear on the cloth, provided that the hole be quite small, but they will appear very confused and imperfect, so much more so as this hole is quite small" ([8], VI, p. 126; [9], p. 97). He further discusses the size of the image being proportional to the object as the focal distance is to the distance between the aperture and the object, and the quality of the image related to the distance to the aperture. His description reflects practical experience with the *camera obscura*. This supports our view that his use of the aperture on the prism was motivated by such experience.

The aperture was later also employed by Newton in the experiment cited above[11]:

> The gradual removal of these suspitions, at length led me to the *Experimentum Crucis*, which was this: I took two boards, and placed one of them close behind the Prisme at the window, so that the light might pass through a small hole, made in it for the purpose, and fall on the other board, which I placed at about 12 feet distance, having first made a small hole in it also, for some of that Incident light to pass through.

Not only where these models for the analysis of the rainbow readily available in *RM*, Descartes also shares the admiration and the principle of wonder with the book. In the last paragraph he writes: "this makes me remember an invention for

[11] *Philosophical Transactions of the Royal Society*, No. 80 (19 Feb. 1671/2), pp. 3075–3087.

making signs appear in the sky, which would cause great wonder in those who were ignorant in the cause". The invention he talks about is an optical experiment described in *RM* and his definition of wonder as "admirables pour ceux qui en ignorent la cause" (repeated from discourse 7) echoes a frequent remark from *RM*. Wonder is an epistemological quality as it depends on our knowledge of the hidden causes. Descartes raises it to a methodological principle.

5 Negotiating Models

Until now we have discussed the successful models as used in Descartes' discovery and explanation of the causes of the rainbow. Evidently, such a sequence of five models, each highlighting some aspect of the phenomenon under investigation, does not comes as a single insight. Each discovery is preceded by a process of careful consideration, trial and error, and experimentation, to find out what the most adequate models may be. Typically for Descartes, we do not find anything of such considerations in his published work. However, the correspondence in the circle of Marin Mersenne provides a witness account of the negotiations on the choice of adequate models for the analysis of specific phenomena.

One interesting example related to color separation is the halo seen around a burning candle (Fig. 8). Leurechon, (*SP*, *Optics*, prop. XX, p. 3) already placed colors around the flame of a candle on the same level as the rainbow. Also in *RM*, the appearance of colors around the flame is compared to the spherical glass filled with water: "or with a candle let the appearances be received upon a shadowed place: you will have the same contentment". Mersenne repeatedly observed a colored corona around the flame of a candle, and discusses the phenomenon in his treatise on the rainbow, written between 1625 and 1627 ([7], II, pp. 649–673). However, from his correspondence it appears that Descartes "is astonished to hear" that Mersenne observed the phenomenon while he himself "rubbed and rolled [his] eyes in all sorts of ways to try to see something similar, but with no success" (Descartes to Mersenne 18 Dec 1629, [8], I, p. 83). A little later Des cartes writes Mersenne that this should not be confused with the corona seen around stars and is to be explained as "secondary light coming from the rays which pass straight through the iris" (Descartes to Mersenne, January 1630, [8], I, p. 106). Descartes' view is confirmed by an experience on a boat in Holland, which

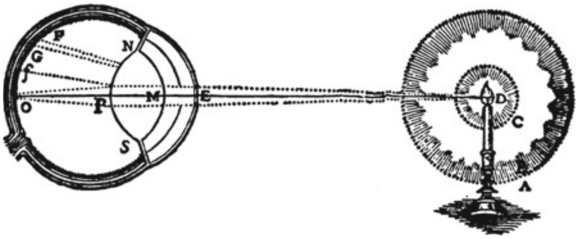


Fig. 8 Descartes' entopical explanation of a halo around a candle

he reported to Golius on 19 May 1635 ([8], I, 318–319): "What this taught me was that the coronas were arranged in exactly the opposite way to those which appear around stars, i.e. red and the outside, and that they did not form in the air, but simply in the water of one of my eyes". Later again, in the ninth discourse of his *Meteoroloy*, he refers to this incident and proposes an explanation of the halos based on the particles on the cornea of the eye ([8] VI, pp. 351–354; [9], pp. 350–351): "their cause must be sought, not in the air, but only in the eye which looks at them". This explanation corresponds with an entry in Beeckman's Journal of 1632.[12] Others like Pierre Gassendi and the Jesuit Claude Dechales instead believed that the phenomenon was caused by the same principles as the rainbow, by dispersion of light. Such differences in opinion cannot be settled in one way or the other. In fact, both explanations have their value. Ocular halos do exist as conjectured by Descartes [26], while halos can also appear around light sources, such as the moon, by dispersion of light through ice crystals. Interesting from a historical standpoint is that natural philosophers were involved in a process of negotiation on the adequate models for the study of specific phenomena. In Descartes' view, the halo around the flame of a candle is a better model for studying entopical diffraction phenomena than for the dispersion of light. Others believed that the flame of candle was a suitable model for studying the causes of color separation. These discussions show us how the selection of adequate models was a process of argumentation and negotiation.

6 Conclusion

The five models that are used in Descartes' analysis are no sophisticated or abstract mathematical models but are rather derived from simple artifacts, daily experiences or technological devices. We have pointed out that (1) the appearance of a rainbow in a fountain, (2) the analogy of a raindrop with a spherical flask filled with water, (3) the kinetic behavior of wooden or ivory balls, (4) the appearance of colors in a prism and (5) the aperture in a camera obscura all appeared as propositions that were the subject of discussion at Jesuit colleges during the disputationes amongst students. Descartes was educated at such a college and was strongly influenced by the Jesuit attempts to physicalize pure and mixed mathematics as part of a pedagogical strategy. He explicitly mentions in his *Principles of Philosophy* that he draws much inspiration from daily artifacts ([8], VIIIA, p. 236, [11]):

> In this matter I was greatly helped by considering artefacts. For I do not recognize any difference between artefacts and natural bodies except that the operations of artefacts are for the most part performed by mechanisms which are large enough to be easily

[12] [6], III, 237: "Iris oculi est humor corneae concavo adhaerens a parte sui pupillae, limbum (ubi lux dissolvens humores, est debilior quam in medio) inficiens, id est tegens. Hinc sequitur iridem circa candelam, aut quodvis lumen pupilla majus, visam, eo videri majorem quo lumen id est ab oculo remotius".

> perceivable by the senses-as indeed must be the case if they are capable of being manufactured by human beings.

Rather than setting up highly contrived experiments, Descartes believes that natural phenomena are better studied by common situations with which we are familiar in daily situations [8], VI, 63, tr. Cottingham): "rather than seeking those which are more unusual and highly contrived, it is better to resort only to those which, presenting themselves spontaneously to our senses". Descartes may have acquired his taste for experience with natural objects and simple artifacts from his days at the Jesuit college of La Flèche (1606–1616). He certainly believed that his approach to physico-mathematics is useful for Jesuit education, as he is recommending his analysis of the rainbow in a letter to Fournet, J.S. of October 1637 ([8], I, p. 455):

> It seems to me that there is no-one who has a greater interest in examining this book than the members of your Society, for I can already see that what it contains (especially as far as meteorology is concerned) will be accepted by so many people that I just do not know how they will be able to teach these subjects without either refuting or accepting what I have written.

Here Descartes seems to be willing to return to the Jesuits what he learned from them: raising the wonder and surprise derived from curious artefacts, contrivances, and machines and employing the practical knowledge embedded in such devices for the purpose of experimentation and analysis in physico-mathematics.

Acknowledgments The author is a research fellow of the Research Foundation Flanders (FWO Vlaanderen) and member of the Centre for History of Science at Ghent University, Belgium. The paper benefited from discussions with Delphine Bellis, Madelina Giurgea and Maarten Van Dyck.

References

1. Armogathe, J.-R.: The rainbow: a privileged epistemological model. In: Gaukroger, S., Schuster, J., Sutton, J. (eds.) Descartes' Natural Philosophy, pp. 249–257. London, Routledge (2000)
2. Boyer, C.: The Rainbow. From Myth to Mathematics. Sagamore Press, New York (1959)
3. Buchwald, J.: Descartes's experimental journey past the prism and through the invisible world to the rainbow. Ann. Sci. **65**(1), 1–46
4. Cusanus.: De ludo globi, 1463, (English translation by Jasper Hopkins), Metaphysical Speculations, vol. 2. The Arthur J. Banning Press, Minneapolis (2000)
5. Dear, P.: Discipline and Experience: The Mathematical Way in the Scientific Revolution. Chicago University Press, Chicago (1995)
6. de Waard, C., (ed.): Journal tenu par Isaac Beeckman de 1604 à 1634. Tome 1: 1604–1619, Martinus Nijhoff, Den Haag (1939)
7. Mersenne, M. Correspondance du P. Marin Mersenne, Religieux Minime, publiée et annotée par Cornelis de Waard; avec la collaboration de Armand Beaulieu; édition entreprise sur l'initiative de madame Paul Tannery et continuée par le C.N.R.S., Editions du Centre national de la recherche scientifique (1932–1988), 18 vols
8. Descartes, R.: Oeuvres de Descartes (11 vols). In: Adam, C., Tannery, P. (eds.) Librairie Philosophique, Paris. J. Vrin (1983)

9. Descartes, R.: Discourse on Method, Optics, Geometry and Meteorology, P. Olscamp (tr.), Hackett, Indianapolis (2001)
10. Descartes, R.: Philosophical Essays and Correspondence. Ariew R. (tr.), Hacket, Indianapolis (2000)
11. Descartes R.: The Philosophical Writings of Descartes, 3 vols., Cottingham, J., R. Stoothoff, D. Murdoch, (tr.) Cambridge University Press, Cambridge (1991)
12. Galison, P.: Descartes's comparisons: from the invisible to the visible. Isis **75**, 311–326 (1984)
13. Giere, R.: Explaining Science: A Cognitive Approach. University of Chicago Press, Chicago (1988)
14. Heeffer, A.: The logic of disguise: descartes' discovery of the Sine law, Historia scientiarum. Int. J. Hist. Sci. Soc. Jpn **16**(2), 144–165 (2006). November
15. Heeffer, A.: Wonder to those that are ignorant in the cause, A critical English edition of Récréations Mathématiques (1624), with a glossary and commentaries. Springer, Heidelberg (2012)
16. Hesse, M.: Models and Analogies in Science. Sheet and Ward, London (1963)
17. Leurechon, J.: Selectae propositiones, Sébastien Carmoisy, Pont-à-Mousson (1622)
18. Mac Lean, J.: De historische ontwikkeling der stootwetten van Aristoteles tot Huyghens. Ph.D. dissertation, Free University of Amsterdam, Van Sijn en zonen, Rotterdam (1959)
19. Magnani, L., Nersessian, N.J., Thagard, P. (eds.): Model-Based Reasoning in Scientific Discovery. Kluwer, New York (1999)
20. Lorenzo Magnani, L., Nersessian, N.J. (eds.): Model Based Reasoning, Science, Technology, Values. Springer, New York (2002)
21. Meheus, J., Nickles, T. (eds.): Models of Discovery and Creativity. Springer, New York (2009)
22. Sabra, A.I.: Theories of Light from Descartes to Newton. Cambridge University Press, Cambridge (1981)
23. Scaino, A.: Trattato del giuoco della palla, Venice: Gabriele Giolito de'Ferrari and brothers, 1555, (Modern editions: a cura di Giorgio Nonni, Urbino Quattroventi 2000. Scaino on tennis, Translated into english by W.W. Kershaw at the instance of C.B. Gabriel, Honorary Secretary and Treasurer of the Royal Tennis Court 1932–1947; with acknowledgements, corrigenda and historical notes by P.A. Negretti. [London]: Strangeways Press, 1951)
24. Shea, W.R.: The Magic of Numbers and Motion. The Scientific Career of René Descartes, Canton (1991)
25. Schuster, J.R.: Physico-mathematics and the Search for Causes in Descartes' Optics—1619–1637. Synthese **185**(3), 467–499 (2012)
26. Simpson, G.C.: Ocular haloes and coronas. Br. J. Ophtalmol. **37**, 450–486 (1953)
27. Westfall, R.S.: The development of Newton's theory of color. Isis **53**, 339–580 (1962)
28. Vaccari, A.: Legitimating the machine: the epistemological foundation of technological metaphor in the natural philosophy of René Descartes. In: Claus, Z., Romano, N. (eds.) Philosophies of Technology: Francis Bacon and His Contemporaries (2 vols.). Bril, Leiden, pp. 287–336
29. Van Dyck, M.: Argumentandi modus huius scientiae maximè proprius'. Guidobaldo's mechanics and the question of mathematical principles. In: Gamba, E., Becchi, A., Bertoloni Meli, D. (eds.) Mathematiche e tecnica da Urbino all'Europa. Max Planck Research Library for the History and Development of Knowledge, Berlin (2013)
30. van Etten, H.: Recreation Mathematique. Jean Appier Hanzelet, Pont-à-Mousson (1624)

Pauli's Idea of the Neutrino: How Models in Physics Allow to Revive Old Ideas for New Purposes

Tjerk Gauderis

Abstract Models have proven themselves to be the key catalyst of many new ideas in science. However, it is not yet fully clarified why models can fulfill such an important heuristic role. The two main reasons stated in the literature—the mental simulation of various scenarios and the wide cross-fertilization across various disciplines—seem to leave out one of the most obvious features of models: they are designed for a purpose. Therefore I investigated why, while the construction of models is a goal-oriented task with a predefined purpose, the use of models yields so many new ideas in science. This paper presents my conceptual analysis together with a detailed historical case study. The functional design of models forces scientists to explore vigorously older ideas to adapt them: as the lacunas in a functional model are also functional, scientists need to modify older ideas (that were formulated for different purposes) to fit the present functional gaps in their models. As such, they construct new ideas. The detailed historical case study exemplifies this by showing how Pauli's original suggestion of the neutrino was, in fact, such an adaptation of Rutherford's earlier idea of the neutron. The present analysis and case study suggest that functional adaptations are salient but often overlooked features of model based investigation.

1 Introduction

Models perform an important heuristic role in scientific investigation ([23, 17]). Their success in this role is typically explained by two widespread practices. On the one hand, because of their dynamic nature, models allow to explore extensively and to experiment mentally with existing theories. As such, one can simulate various scenarios, and identify and mediate lacunas and anomalies in a given

T. Gauderis (✉)
Centre for Logic and Philosophy of Science, Ghent University, Gent, Belgium
e-mail: tjerk.gauderis@ugent.be

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_24, © Springer-Verlag Berlin Heidelberg 2014

theory [17, 18]. On the other hand, models can be extremely fruitful. Many simple abstract models have been applied to a variety of contexts outside their original field, a process that has led to a huge interdisciplinary cross-pollination.[1] This shows that scientists are actively looking for useful models that can be applied to problems in their own field.

In this paper, I want to add a third important heuristic practice involving models. Models have a typical functional structure or, to put it in other words, they are designed with a purpose in mind. This means that lacunas or gaps in a model are by definition also functional, which invites the designer to actively explore old ideas that might serve to fill these gaps.

The aim of this paper is to explicate and illustrate this third practice and its relation to the other two by presenting a case study of a famous idea in the history of modern physics, i.e. Pauli's original suggestion in 1930 of the particle that was named the *neutrino* afterwards. In Gauderis [11], I have argued that this idea was not so original as Pauli might have thought it to be, but can be seen as an adaptation of Rutherford's original idea of the *neutron* in 1920. In this paper, I will further substantiate this claim by explicating both the problems that Rutherford and Pauli were working on, as well as the models they employed for their purposes. This analysis will interpret this history as an example of how an idea that arose in a certain program managed to stay alive, even though the program grew obsolete, in order to be picked up ten years later as the missing piece in a model for another and much more prominent puzzle in the field.

In the next section, I will start by expanding on the role of models in scientific discovery, and show how the three heuristic practices identified above can be understood in generic terms, although the scope of this analysis will be restricted to the heuristic use of models in physics. The main objective of this section is to provide us with a conceptual framework to analyze the case study, which naturally falls apart into two distinct parts: Rutherford's reasoning and models in 1920 and Pauli's reasoning in 1930, each of them being discussed in a separate section.

2 Models and Scientific Discovery in Physics

As some scholars have noticed, it is very hard to give a precise definition of a model, even if we restrict ourselves to models in physics [13, p. 52; 18, p. 12]. Such a definition is, however, not necessary and one can content oneself, as

[1] There are numerous examples of this cross-fertilization. Some of the more spectacular examples are the so-called genetic algorithms in Artifchain-models to identify authors inicial Intelligence, which are based on natural evolution models (e.g. [12]), the use of Markov philological studies (e.g. [16]) and the use of phase transition models from physics for problems in social philosophy of science (e.g. [8]).

Nersessian proposes, with a loose definition that is sufficient to capture the way physicists think of models.[2]

From the most general point of view, a model can be conceived as an abstract imaginary system of interrelated parts that has as a whole certain distinguishing characteristics. The most important characteristics of models are (1) their functional design, (2) their representational potential and (3) their susceptibility to manipulation, all of which are uncontested in the relevant literature. Models in science are in the first place functional, or as Morrison and Morgan state it, they are designed or constructed to "function as tools or instruments" [17, p. 11]. The purpose of this design can be a variety of scientific activities such as theory construction, explanation, prediction, suggesting which data should be collected, etc.[3] The main reason why models can function as a tool for all these purposes is their second property: they are meant to represent certain features of the world, or as Nersessian explains it, "they are designed to be structural, functional or behavioral analogues of their target phenomena" [18, p. 12]. Finally, the reason why models can be considered as tools is their susceptibility to manipulation, and as such they distinguish themselves from mere representative descriptions. The design of a model is such that one can interact mentally with the model by manipulating certain features, adjusting certain parameters or adding or removing certain parts, all of which represent interventions in the target field [18, p. 12].[4]

With this characterization in mind, we can explain how models play their role in the three heuristic activities identified in the introduction. When a scientist is confronted with a new target phenomenon, i.e. a collection of experimental data, she can try to structure this data by constructing a model. She does not have to start this activity from scratch. Generally, some initial constraints are available from a general theory or some related models. For example, if a researcher tries to construct a model for a particular type of nuclear reactions, initial constraints are raised by her model of nuclear constitution and some general theories such as quantum mechanics. Still, it is obvious that there are no clear algorithms at this stage, but that this is more a matter of skill. In the literature, this activity has been described as constructing with bits from different sources [18, pp. 15–16], as

[2] By purely focusing on the epistemological role of models, we evade at the same time the discussion about the ontology of models. For an introduction to the latter discussion, see Frigg and Hartmann [10].

[3] Epstein [9] distinguished seventeen different reasons why one should model. Apart from the most straightforward reasons such as prediction and explanation, he identifies models also as a key method to e.g. finding gaps in your data or formulating new research questions. While models are often constructed for several of these reasons, he convincingly argues that e.g. models for prediction and explanation are typically of a different nature. Not all of his seventeen reasons, however, should be considered as purposes for the design of a particular model. His paper aims mostly to convince scientists to use models in fields where they are less common, and some of his reasons such as "teaching us a scientific outlook" are just interesting qualities that result from the regular use of models.

[4] Because of this feature, the use of models has received a central place in the interventionist view of science going back to Hacking and Cartwright [5].

matching representations to mathematical structures [7] or as constructing a hybrid between target and source domain [18]. The common denominator is that this activity is a bit-piece assembly process. This assumes that scientists have certain simple blueprints at hand, simple mathematical models and structures that they have acquired over the years, and which they can be combine with theory, experimental data and various representations. The so-called model constructing skills consist then in maintaining and expanding such a set of blueprints, and the application of this set to various problems is exactly the second heuristic practice that I have identified.

Second, the models' susceptibility for manipulation allows us to simulate and explore within the constraints that are imposed on the model by both its internal (formal) coherency and the knowledge of the target domain. This hybrid construction gives the models a relative autonomy, which allows them to identify lacunas and anomalies in both the theory and the model itself, the first heuristic practice stated above.

Now we can specify the third practice that makes models such a useful heuristic tool. The functional design of a model ensures that every part of the model has its own function. As such, if a lacuna is identified in a model, this is a functional lacuna, i.e. the model misses something that can fulfill a function needed by the rest of the model. Researchers can try to come up with own and original ideas to fill these gaps, but, as the case study exemplifies, it is also very common that a researcher browses her own field for ideas that have the necessary properties. This is a different activity than the use of various models from other fields. Where in the latter activity the abstract structure of the model is borrowed and completed by adjusting the representational elements to objects in the field the researcher is working in, in the former activity, the researcher actively pursues ideas from her own field that were proposed for different purposes or problems, some of which might have already become obsolete.

To sum up, I have identified three scientific heuristic practices involving models, which explain why the use of model is heuristically so successful. First, models allow by their partly formal nature to be applied to many problems that can be situated in other fields, or shed at least some initial light on these problems. Second, they allow for dynamic simulation on the basis of which researchers can explore the various combinations of the model's parameters. Third, because of their functional design they invite scientists to actively reconsider old ideas in order to spot an idea that has the right characteristics to fulfill a particular function in the grand design of the model.

In order to apply these concepts to the case study, I will first expand on the two types of models that will be discussed in the case study, i.e. constitution models and reaction/process models. *Constitution models* are the oldest type of physical models and relate to ancient philosophical questions about the nature of things. The main purposes of these models are explanatory: by specifying the various parts and the total structure of the target phenomenon, one aims to explain certain properties of the whole, such as its stability or fluidity. As *process* and *reaction models* started to emerge only since the scientific revolution, they are much younger types of models. Their main purpose was not so much to explain the nature of the represented changes, but rather to explicate the necessary conditions and the results to be expected. Experimentalists used them as a guide line to manipulate and control physical reality. In other words, these types of models are designed for prediction and not for explanation. The scientific revolution and these experimental models put at the same time more stringent conditions on the older constitution models: they had to become compatible with (most of) the process and reaction models related to the target phenomenon and their descriptions had to be limited to qualities that are in principle testable. Therefore, constitution models in physics are generally limited to describing the various subentities and specifying the forces or mechanistic properties that keep them together. Still, their main purpose remains explanatory as they are not strictly needed for prediction.

Both types of models are—as all models are—dynamic by nature and allow the scientist to interact with their various parts. For constitution models, these dynamics lie mostly in the possibility to explore what combinations of subentities can possibly exist. To discuss the dynamics, I need to distinguish between process and reaction models. I view reaction models as models that take the represented change to occur instantaneously, e.g. models for radioactive decay or chemical reactions. In contrast with process models, which represent a gradual or stage-based change of the target phenomenon, reaction models represent only the situation before and afterwards, considering the change as something that has happened at a certain time in between. The main goal of these models is to specify, apart from the conditions under which the reaction can take place, which characteristics and entities are conserved and how the non-conserved properties change. Their main dynamics lies in the fact that one can mentally explore various situations to picture what the result of the reaction would be. Process models, which do not occur in the case study, draw the attention more to the change of the target phenomenon itself, and enable scientists to simulate various scenarios how they might control or accommodate the process.

By ascribing different purposes to these different types of models, while still assuming their compatibility, I take models to be more or less autonomous but related to each other. The autonomy of the models' purposes is also one of the reasons why Morgan and Morrison [17] consider models to be independent from physical theories. The other reason is that models are also constructed more or less autonomously from theory or as Cartwright [4] states it: "Theories do not provide us with algorithms for the construction of models, they are not vending machines into which one can insert a problem and a model pops out". This relative autonomy will also help us to understand the role of theories such as quantum mechanics and classical electrodynamics in the case study. The semantic view (in which theories are superfluous families of models) and syntactic view (in which the logically structured theory carries all scientific value) are both too restricted to capture how theories and models function in the endeavors of scientists.[5]

[5] See Frigg and Hartmann [10] for an excellent summary of these two points of view.

Cartwright [3] has described the laws in a theory as "schemata that need to be concretized and filled in with the details of a specific situation, which is a task that is accomplished by a model", and such initially under schemata constructed models can develop their own dynamics, which might lead to the suggestion to withdraw a certain aspect of the theory (see e.g. Bohr's suggestion to withdraw the energy conservation theorem [11]).

3 Introduction to the Case Study

The case study in this article is Pauli's original suggestion of a particle that was called the neutrino by Fermi in 1933. In Gauderis [11], in which I discussed the various proposed hypotheses to solve the anomalous β-spectrum in the late 1920s, I have argued that Pauli's idea of the neutrino was basically an adaptation of the old idea of the neutron suggested by Rutherford in 1920. In this article I take on the challenge to explicate how this adaptation should be understood in terms of the given characteristics of models. The case study naturally falls apart in two parts. I will first take the time to explain Rutherford's project around 1920 and show how the idea of the neutron emerged from his model. Next, I will present the other case, which was the completely different puzzle of the curious β-spectrum. I will restrict myself however to Pauli's suggestion, and show how the model he had in mind led him to think that the neutron might be the solution.

4 Rutherford's Idea of the Neutron

Rutherford suggested the idea of the neutron for the first time in his Bakerian lecture [24]. The main reason why he believed this idea to be valuable was because he thought that its existence "seems almost necessary to explain the building up of the nuclei of heavy elements" (p. 397). Translated to our conceptual framework, Rutherford perceived an incompatibility between his constitution model of atomic nuclei and the theory of classical electromagnetism, because the laws of the latter do not allow the building up of the more heavy nuclei. This had led him to investigate further the constitution of nuclei, partly by real-life experiment, partly by mental simulation of the model and logical thinking. It was exactly this simulation of various possibilities that convinced him that there might possibly exist a neutron, although his perception of it was totally different than our current understanding. The fact that this idea, yielded by simulation of the model, could fill the functional gap in the model convinced him of the soundness of this idea, a conviction that inspired him to look tenaciously for experimental proof over the next ten years. Finally, in 1932, his close collaborator Chadwick managed to assemble sufficient evidence to confirm its official discovery.

Let us first explain Rutherford's nuclear model. Like many of his contemporaries, he believed that the nucleus consisted out of "electrons and positively charged bodies" such as helium and hydrogen nuclei (p. 377). But, as he had already suggested in 1914, all these positively charged bodies can ultimately be considered as a combination of positively charged hydrogen or H-nuclei (which became gradually called *positive electrons* or *protons*) and negatively charged *electrons*,[6] kept together by the electromagnetic force[7] [19, p. 230; 24, p. 395]. For example, the He-nucleus or *α-particle* was considered to be a very stable combination of four H-nuclei and two electrons. This so-called *proton-electron* or *p-e model* explained convincingly the atomic mass and charge of the various elements. A nucleus with atomic mass *A* and charge *Z* consisted of *A* protons (which all have, as they are hydrogen nuclei, elementary mass, and, hence, add up to the atomic mass *A*) and *A* minus *Z* electrons (which made sure that the total charge of the nucleus was positive *Z*). The Coulomb force between these positive and negative particles caused the nucleus' stability.[8]

Given the common ontology of these days, this was the only viable model available. Still, this model had several difficulties.[9] As Rutherford mentions, the apparent lack of magnetic moment of the intranuclear electrons hints that these electrons must be somewhat "deformed" (p. 378) and that they are in no sense comparable to the extranuclear electrons orbiting around the nucleus. But his main problem was the constitution of large nuclei. As soon as a nucleus contained a certain amount of protons, the combined repelling Coulomb force of these would be just too large to ever let another proton come close enough to swallow it. The reason why he was so vividly aware of this problem, was because he observed it on a daily basis in his experiments. While he found it possible to shoot lighter elements with α-particles or He-nuclei, and initiate a collision, he found it impossible to penetrate larger nuclei due to their high electrostatic repulsive forces.[10]

[6] Hanson [14, pp. 157–159] explains the fact that for a long time scientists refused to consider any other elementary particle besides protons and electrons by pointing to the fact that these two particles were at the same time considered to be the elementary subunits of the two types of electrical charge. As there was no other type of electricity, there was no reason to presuppose another elementary particle.

[7] The only two forces known at the time were gravity and electromagnetism, but, because gravity is too weak to play a role at such a small scale, the only viable option was electromagnetism.

[8] Notice the contrast with our present day views, in which we take a nucleus to consist of *Z* protons (accounting for the nuclear charge) and *A* minus *Z* neutrons (adding the total mass up to *A*), kept together by the residual strong force. For example, it is now thought that the He-nucleus consists of 2 protons and 2 neutrons instead of 4 protons and 2 electrons.

[9] At this point, I only mention problems that were already known in 1920. The more famous problems for this *p*-e model, such as the wrong statistics of the nitrogen nucleus and the Klein paradox, arose only during the 1920s.

[10] It was exactly because part of the α-particles were repelled from the gold foil in the famous Rutherford-Mardsen-Geiger experiments that Rutherford inferred the existence of the nucleus in the first place.

This also nicely illustrates that Rutherford perceived experimental data, models and theories all as more or less autonomous entities that should be made compatible with each other.

Rutherford thought he could cope with these problems by assuming certain substructures in the nucleus. Nuclei were not just a heap of protons and electrons that attract all of each other more or less equally. He thought that protons and electrons bound in small stable substructures, which in turn grouped together to form the full nucleus. The reason why he (and the physics community in general) had this idea was the remarkable stability of the α-particle. In experiments it turned out to be impossible to break up this element by collisions [24, p. 379]. It was also observed as an independent structure in α-decay, which led several people to assume that it was as such part of the nucleus. Around 1920, it was Rutherford's main experimental program to find more stable combinations like this to complete the nuclear constitution model. Because he was not able to reach the nucleus of heavier elements with α-particles, he conducted mainly experiments on lighter elements (nitrogen, oxygen, carbon) in order to produce collisions and study the remaining parts. His first discovery were H-nuclei or protons. This was important because, although it was generally assumed that protons existed independently in the nucleus, it was "the first time that evidence has been obtained that hydrogen is one of the components of the nitrogen nucleus." (p. 385), and that, hence, the *p-e* model had some experimental ground. Second, he discovered a certain atom, which he called X, with atomic mass 3 and nuclear charge 2, which made it "reasonable to suppose that atoms of mass 3 are constituents of the structure of the nuclei of the atoms of both oxygen and nitrogen" (p. 391).

In order to figure out the substructure of this atom X, he reasoned that "from the analogy with the He-nucleus, we may expect the nucleus of the new atom to consist of three H-nuclei and one electron." (p. 396), which made this atom a snug fit in the *p-e* model. But when he realized that this means that a single intranuclear electron can bind three protons,[11] it appeared to him "very likely that one electron can also bind two H-nuclei and possibly also one H-nucleus" (p. 396). In other words, by mentally exploring what is also reasonable to expect according to the *p-e* model, he came to the idea of a close binding of one proton and one electron, an "atom of mass 1 and zero nucleus charge". He expected this combination, which he started to call the neutron later, to be a very stable entity with "very novel properties". Because there would hardly be any electromagnetic field associated with this neutral combination, it would be able to travel rather freely through matter. Therefore, it might reach the nucleus of heavy atoms without suffering from a repelling force, where "it may either unite with the nucleus or be disintegrated by its intense field" (p. 396).

[11] A single intranuclear electron was not yet observed beforehand, hydrogen had according to the *p-e* model no intranuclear electrons, while the next element in the periodic table, helium, had already two.

The thought process of how Rutherford came to the idea of the neutron is highly intriguing, because hardly any part of it is still acceptable according to our present standards: the *p*-*e* model is plainly wrong; later experiments did not confirm the existence of the X-atom; the whole idea that there exist certain substructures in the nucleus is flawed; the constitution of heavy nuclei poses no problems; and above all, according to our present understanding, it is absolutely untrue to consider a neutron as a combination of a proton and an electron.[12] Still, judged in light of Rutherford's background knowledge, his thought process is a very sane and sound piece of reasoning in which he improved his constitution model by combining experimental data with mental simulation of his model. And, although Pais claims that this whole search program for atomic substructure has left no mark on physics [19, p. 231], I have shown that this program has led to a valuable idea, which is not only the forerunner of our current neutron, but also, as I will show in the next section, of our current neutrino.

5 Pauli's Idea of the Neutrino

Pauli's suggestion was an attempt to answer a very complex experimental puzzle, which had been intriguing the physics community for several years. In 1927, Ellis and Wooster had published an experiment that convincingly showed that the electrons in radioactive β-decay were emitted with a broad and continuous range of energies. This puzzling fact did not only break the analogy with α-decay, in which the energy of the emitted α-particles was determined for each possible α-decay, it also triggered some very counterintuitive hypotheses. For example, Rutherford and Chadwick suggested that not all nuclei of a particular β-unstable element were identical, because they had different internal energies, and Bohr suggested that energy is not conserved in β-decay such that the electrons can escape with a wide range of different energies.[13] At the time, this puzzle was not the only problem for nuclear research. The experiments performed in 1926 that showed that nitrogen nuclei behaved according to Bose–Einstein statistics, proved another serious anomaly for nuclear theory and the *p*-*e* model. According to latter, nitrogen (with atomic mass 14 and nuclear charge 7) consisted out of 14 protons and 7 electrons, and should, hence, have a half-integer total spin, because both protons and electrons have spin ½. This means that, according to the *p*-*e* model, nitrogen nuclei should behave according to Fermi–Dirac statistics. The observed Bose–Einstein statistics, however, required that the nucleus consisted of an even

[12] The idea that neutrons were not close combinations of protons and electrons took some time to settle. When Heisenberg published in 1932 the first elaborate proton-neutron model of the nucleus, he left the question still open [1].

[13] For an extensive exploration of this puzzle and all suggested hypotheses, see Gauderis [11].

number of half-integer particles, adding up to an integer total, which is required to explain these statistics.

Pauli was introduced to these problems in 1929 by Bohr, who was thinking about a restriction of the principle of energy conservation to solve these issues, an idea that gave Pauli "very little satisfaction" (according to a letter to Bohr reprinted in [22]). As Bohr and his collaborators continued this path of energy non-conservation, Pauli started, apart from continuously criticizing them, thinking about another idea, which he formulated for the first time in December 1930. Let us first, as we did with Rutherford, try to understand Pauli's view on and models of the matter. We can then explain how he adapted Rutherford's idea for his own purposes.

Like all of his contemporaries, Pauli saw radioactive decay in terms of a reaction model in which an unstable nucleus (the situation before) decayed spontaneously into a remnant nucleus and the observed emitted α- or β-particle plus some γ-radiation (the situation afterwards). For α-particles, the model of this reaction preserves both energy and electric charge, but for β-decay, the unexplainable continuity of energies in the situation afterwards had led some to suppose that this continuity already existed in the situation beforehand (Rutherford and Chadwick [25]), or to suggest to retract the energy conservation constraint for this model (Bohr). It was this final suggestion that triggered Pauli to address this problem. To understand why he was so opposed to Bohr's ideas, we have to look at some of his criticisms. In a letter to Klein, a close collaborator of Bohr, he challenges Bohr's suggestion to retain charge conservation but abandon energy conservation in β-decay by the following thought experiment:

> Imagine a closed box in which there is radioactive β-decay.[…] If the energy law thus would not be valid for β-decay, the total weight of the closed box would consequently change. This is in utter opposition to my sense of physics! For then it has to be assumed that even the gravitational field – which is itself generated […] by the entire box (including the radioactive content) – might change, whereas the electrostatic field, […], should remain unchanged because of the conservation of charge (reprinted in [15]).

The heart of Pauli's criticism is that the field formalisms for gravity and electrostatics, both depending on inverse-square laws, are constructed analogously and, hence, considered to be of the same kind. By breaking this analogy, Bohr's suggestion has the far reaching consequence of undermining the physical concept of a field. Unlike most quantum theoreticians who had hardly ever to deal with gravity,[14] Pauli was also an expert in the field of general relativity.[15] Because of this, Pauli was much more aware of field structures as the main ontological concepts for physical reality than nuclear physicists in general. This explains why Bohr's ideas were so disturbing for him.

[14] The effects of gravity are far too weak to be noticeable at the atomic scale.

[15] At the age of 21, he wrote a summarizing monograph on the general theory of relativity, which impressed even Einstein [20, p. 215].

If Pauli was convinced that the conservation laws must hold in this reaction model, it probably occurred quickly to him that the only way to balance the disequilibrium between the before and after situation was by adding something to the picture. But nothing else was observed so far in the β-decay experiments. Hence, he needed to look for something that was unobservable or at least very hard to observe. As conservation of electrical charge already applied, it had to be also electrically neutral. In other words, his reaction model for the process of β-decay had suddenly a gap, which should be filled by an idea or entity which had these two properties. In an autobiographical article, he wrote the following:

> Then I have tried to connect the problem of the spin and statistics of the nucleus with the other problem of the continuous β-spectrum without giving up the energy conservation principle through the idea of a neutral particle. I have sent a letter about this [...] in December 1930, when the heavy neutron was not yet experimentally found [21, p. 1316, my translation].

In this letter[16] he presented a "desperate remedy" to solve these two problems, namely "that there could exist electrically neutral particles in the nucleus, which I want to call neutrons" [21, p. 1316, my translation]. In Gauderis [11], I have argued that this idea was largely an adaptation of Rutherford's idea. The main arguments for this thesis are, first of all, the fact that Pauli used the term "neutron" for this nuclear constituent, while the term "neutron" was still actively used by Rutherford and his collaborators in articles, a fact Pauli must have been aware of. Second, the fact that he points out why Rutherford had no success in finding his neutron before 1930 right before presenting his own idea hints that he thought to have found what Rutherford was looking for. Finally, he seems to have abandoned the idea that his particle was a nuclear constituent only around 1932, the year in which the "heavy" neutron was discovered by Chadwick [6].

The functional gap in Pauli's model was clear: he needed something that could carry some spin and energy, had zero charge and was very hard to detect. At the same time, there was an old and well-known hypothesis of a neutral particle that is hard to detect. Although it had been suggested in a completely different context and was aimed at another purpose, its properties made it suddenly a viable candidate to fulfill the functional gap of another problem. As such, Pauli employed this idea in his model, calculated its further properties such as its spin and possible mass, and was able to put forward the first version of his attempt to solve the β-puzzle.

6 Aftermath of the Case Study

In 1932 Chadwick, a close collaborator of Rutherford at the Cavendish laboratory, announced the experimental discovery of the neutron. In his article [6] he stated explicitly that what he found was the particle Rutherford envisioned in his

[16] The original German text of this letter can be found in Pauli [21, p. 1316]. An integral translation can be found in Brown [2, p. 27].

Bakerian Lecture in 1920 [24]. This discovery was directly accepted by the physics community, and gave rise to the first proton-neutron models of the nucleus later that year, which were able to explain the anomalous statistics of nitrogen nuclei. Pauli's hypothesized particle, which had, in contrast with the discovered "neutron", a very low mass, was dubbed the "neutrino" by Fermi in 1933 to distinguish it from Chadwick's discovery. Gradually, this solution for the anomalous β-spectrum drew more adherents until it was incorporated in Fermi's model for β-decay in 1934, after which consensus followed shortly after. Bohr himself admitted defeat in 1936. However, it took until 1956 before experimental evidence was found for the neutrino.

7 Conclusion

In explaining the success of using models for heuristic purposes, the functional design of models is often left out of the picture. In this paper, I showed by means of a conceptual analysis and a detailed historical case that precisely this functional design of models forces researchers to explore vigorously old ideas in order to adapt them for their current purposes. As such I identified a third practice—besides the often mentioned mental simulation of various scenarios and wide cross-fertilization between different fields—that explains the heuristic success of models.

Old ideas are often reused, generally adapted or employed as an analogy or metaphor. The case study in this paper explains in detail how Pauli filled the functional gap in his model for radioactive β-decay by adapting an old idea that figured in Rutherford's atomic constitution model. But not only entities or objects serve as ideas that can be adapted for new purposes; this is illustrated by Bohr's suggestion to retract the energy conservation principle. At least twice before did he use this same idea to solve a certain puzzle, each time with a completely different purpose.

Although it is for many cases impossible without a detailed case study to tell whether it concerns the use of an old idea or whether one came independently to the same idea, there is certainly no reason to suspect that all these ideas were original. As such, if we want to understand how scientists use models and reuse them, it is important to be aware of how models invite scientists by their functional structure to actively explore old ideas in order to adapt them for their own purposes. Further case studies and formal analyses are, however, needed to understand the impact of this fact on our current methodologies.

References

1. Bromberg, J.: The impact of the neutron: Bohr and Heisenberg. Hist. Stud. Phys. Sci. **3**, 307–341 (1971)
2. Brown, L.M.: The idea of the neutrino. Phys. Today **31**(9), 23–28 (1978)

3. Cartwright, N.: How the Laws of Physics Lie. Oxford University Press, Oxford (1983)
4. Cartwright, N.: The Dappled World. A Study of the Boundaries of Science. Cambridge University Press, Cambridge (1999)
5. Cartwright, N., Shomar, T., Suárez, M.: The tool-box of science. In: Herfel et al. (eds.), Theories and Models in Scientific Processes, pp. 137–150. Rodopi, Atlanta (1995)
6. Chadwick, J.: The Existence of a Neutron. Proc. R. Soc. **A136**(830), 692–708 (1932)
7. Czarnocka, M.: Models and symbolic nature of knowledge. In: Herfel et al. (eds.) Theories and Models in Scientific Processes, pp. 27–36. Rodopi, Atlanta (1995)
8. De Langhe, R.: To Specialize or to Innovate? An Internalist Account of Pluralistic Ignorance, Synthese (2012)
9. Epstein, S.: Why model?. http://www.santafe.edu/media/workingpapers/08-09-040.pdf (2008)
10. Frigg, R., Hartmann, S.: Models in science. In: Zalta, E. (ed.), The Stanford Encyclopedia of Philosophy (Fall 2012 Edition). http://plato.stanford.edu/archives/fall2012/entries/models-science/ (2012)
11. Gauderis, T.: To envision a new particle or change an existing law? Hypothesis Formation and Anomaly Resolution for the curious Spectrum of β-Decay (2013)
12. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Boston (1989)
13. Hartmann, S.: Models as a tool for theory construction: some strategies from preliminary physics. In: Herfel et al. (eds.) Theories and Models in Scientific Processes, pp. 49–67. Rodopi, Atlanta (1995)
14. Hanson, N.R.: The Concept of the Positron. Cambridge University Press, Cambridge (1963)
15. Jensen, C.: Controversy and Consensus: Nuclear Beta Decay 1911–1934. Birkhäuser Verlag, Basel (2000)
16. Khmelev, D.V.: Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts. J. Quant Linguis. **7**(3), 201–207 (2000)
17. Morgan, M., Morrison, M.: Models as Mediators. Cambridge University Press, Cambridge (1999)
18. Nersessian, N.: Creating Scientific Concepts. MIT Press, Cambridge (2008)
19. Pais, A.: Inward Bound of Matter and Forces in the Physical World. Oxford University Press, Oxford (1986)
20. Pais, A.: The Genius of Science. Oxford University Press, Oxford (2000)
21. Pauli, W.: Zur älteren und neureren Geschichte des Neutrinos. In: Kronig, R.,Weisskopf, V. (eds.) Wolfgang Pauli. Collected Scientific Papers (1964, pp. 1313–1337). Interscience Publishers, New York (1957)
22. Peierls, R.: Introduction to Peierls R. (Ed.), Niels Bohr Collected Works, vol. 9. Nuclear Physics (1929–1952) (pp. 3–84). North Holland Physics Publishing, Amsterdam (1986)
23. Redhead, M.: Models in Physics, Brit. J. Philos. Sci. 31: 145–163 (1980)
24. Rutherford, E.: Bakerian lecture: nuclear constitution of atoms. Proc. R. Soc. **A97**(686), 374–400 (1920)
25. Rutherford, E., Chadwick, J.: Energy relations in artificial disintegration. Proc. Cambridge Philos. Soc. **25**, 186–192 (1929)

A Historical Case Study of a “Nested Analogy”

Nora Alejandrina Schwartz

Abstract In order to understand human neuromuscular function, Luigi Galvani had to face the problem relative to the way in which “animal electricity” may be stored in animal tissue. Searching for a solution, Galvani built a hybrid model which constitutes a “nested analogy”. This model satisfies constraints provided by objects of his work environment. I defend the claim that the model visual structure is based on a factual or physical scenario—as opposed to counterfactual or imaginary. But this assertion does not imply that the visual structure comes from an existing structure recognized as the target.

Nancy Nersessian has shown that during scientific innovation episodes, models to reason about real world problems under investigation are usually created. She has exhibited that hybrid representations used as sources of analogical thinking in science are a usual case of those models. These representations are hybrid as they are intermediary buildings that satisfy constraints of the target and the source domains [1]. Chandrasekharan and Nersessian have pointed out that hybrid models can constitute nested analogies. These authors have also established that current theories of analogy do not consider, among others, a feature of built analogies: *that the visual structure of the model is not based on an existing structure recognized as the target*. Using a case study, they concluded that the visual structure of the model at issue is based on an imagined structure [2]. In this paper I will analyze the “second order” hybrid analogical model which Luigi Galvani built in the process named *animal electricity discovery*. I agree with Chandrasekharan and Nersessian in their negative claim mentioned above. But I will argue that in Galvani’s case the visual structure of the analogy source comes from a factual scenario and not from a counterfactual one.

N. A. Schwartz (✉)
Economics Faculty, University of Buenos Aires, Buenos Aires, Argentina
e-mail: nora_schwartz@yahoo.com.ar

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_25, © Springer-Verlag Berlin Heidelberg 2014

1 The Animal Leyden Jar: A Hybrid Analogical Model of "Second Order"

Chandrasekharan and Nersessian use the expression "nested analogy" meaning the case in which a model provides an analogy to solve a problem relating to a system that, on its own, is an analog model of a real system. In this way, the "animal Leyden jar" is an analogical source from which inferences are made about animal models that, on their own, are analogs of human beings [2]. From that *constructo*, Galvani tried to determine in which place of the frog the animal intrinsic electricity may be stored and, more generally, to establish how an unbalance of opposite electrical charges can exist within an animal organism.

Galvani was a physician interested in the therapeutic use of electricity. This was a practice that began to be exercised near 1744. By then the expression "Medical Electricity" made reference to electric shocks and spark applications in order to manage various illnesses, in particular palsy and "nervous disorders" [3]. Because of its controversial efficiency, scholars noted the need of a deeper understanding of the physiological mechanisms that control bodily functions in the application of electricity [4].[1]This was one of the factors that made Galvani direct his attention to neurophysiology [5].

During the XVIII century the possible implication of electricity in the nervous function and in muscle excitability was a main theme of interest. There were two views which intended to explain the muscle movement: the theory of irritability and the neuroelectric theory.[2] The presupposed theory of electricity (by both of the

[1] "Medical Electricity" arrived to Italy in 1745. Luigi Sale was mentioned among the first ones to use it as a curative therapy. In Venice, Gianfrancesco Pivati, pretended having discovered a new method in the application of electricity to cure deeply rooted illnesses immediately. Since the publication of his book *Dell'elettricità medica* in 1747, it arose a controversy about the use of electricity in medical therapy in Italy. It lasted until 1750 and expanded over all Europe. Because of that, the Academy of Bologna asked Giuseppe Veratti the task of testing experimentally the efficiency of the therapeutic method proposed by Pivati. The experiments -which confirmed the goodness of the method—were published in the book *Osservacioni fisico-mediche intorno all'elettricità* in Bologna during 1748. However, Veratti was careful in accepting Electric Medicine without any restriction. Galvani took this point of view, too.

[2] According to Albrecht von Haller, muscles have an intrinsic capacity ("irritability") to contract in response to physiological or experimental stimulations. For instance, in response to a needle prick. Nerves may produce muscle contraction activating this intrinsic capacity, but they are not effective agents of the contraction. On the other hand, the neuroelectric theory, held by Tommaso Laghi, defended that the matter contained within the nerves, presumably of electric nature, is the "efficient" cause of the muscle movement.

mentioned views) was Franklin's [6].[3] The confrontation between hallerians, supporters of the theory of irritability, and the partisans of the neuroelectric theory[4] helped to define the conditions that might be met by any electric theory of the nervous conduction in order to be physically and physiologically plausible. Explaining how electrical unbalance—necessary to move the electric fluid along nervous fibers—could be within an animal organism, in spite of the conductive nature of the body tissues, was thought a main requirement. More precisely, discussion between members of each side showed the need to answer the following question: if nerves (like muscles and surrounding tissues) were conductors, there could not exist any electrical unbalance in animal organism, because conductor humors are able to dissipate any electrical unbalance generated within them. And so, no muscle contraction could be produced. So it must be explained how electrical unbalance can exist within animal organism, in spite of the conductive nature of the body tissues [7].

Galvani thought that studying *animal* anatomy and physiology was essential to understand in which way electricity is involved in *human* neuromuscular function. He developed his research through series of well planned experiments with animals. "Animal model" is an expression recently coined, but it can be used properly to call the animals with which Galvani made experiments. W. Burggren says that "(…) an animal model is an animal that is studied frequently and preferentially to tell us more about systems, tissues, cells and biochemical and physiological processes in which we are interested" [8]. "Amphibians have many compelling features that make them ideal as animal models (…): basic physiology, diversity, favorable phylogenies, wide range of habitats, temperature and oxygen tolerances, sufficient similarities to mammals, and straightforward maintenance" [9]. In *De Viribus* Galvani said that his purpose was to transfer relevant information about animals, specially cold blood animals, to human beings. In fact, Galvani decided to study the physiological role of electricity in a *"prepared" frog*.

[3] Franklin's theory of electricity held that there is only one sort of electrical fluid which exists in all bodies. The body that contains a normal amount of electricity or which is in balance is a "non electrified" body; this implies that it produces non observable effects. However, a non electrified body can become unbalanced: it can gain electrical fluid and thus, reach a positive state (*plus*), or it can lose something of its natural amount, resting in a negative state (*minus*). Electrification is the process by which a body with a normal amount of electricity receives further fluid from a positively electrified body and gives fluid to a negatively charged body. Electrification can be produced through conduction—like when two bodies touch each other or are near enough to allow a spark to pass between both of them through the air. A conductor can make the electricity flux move by making contact between opposed electricity charges, thus, restoring equilibrium in an unbalanced body.

[4] The debate among hallerian and neuroelectric theories supporters was particularly strong in the 1755–1760 period. In Bologna, Italy, Fontana was on the side of Haller, and the academic establishment members defended the electric nature of the nervous conduction. Since 1750 to 1770 the hallerian system became accepted by the majority. But, this situation would change. Since 1770 the neuroelectric theory was rehabilitated as a consequence of research on "electric fish", which showed that some animals, like the *Torpedo* and the *Gymnotus* have inherent electricity by nature [5].

He prepared the frog like the experimental researchers who followed Haller. The part to be examined was completely isolated in order to work on it. The frog was decapitated, its crural nerves were uncovered and an electrified rod was brought near them [10].

On developing his research,[5] the observation of a nervous fluid circuit in the frog, the pattern of which was like the perceptual structure of a Leyden Jar electrical discharge, made him infer that the nervous fluid is an *electrical current of discharge,* in this case of a conjectured frog inherent electricity [11]. He seemed to assume that there is an intrinsic animal electricity stored that, when unbalanced, is discharged across that path shape. So, he wondered in which place of the frog—similar to that physical condenser—the animal intrinsic electricity might be stored. This question involved solving the hallerian objection mentioned above: how could an unbalance of opposite electrical charges exist within an animal organism? In the process of searching an explanation for this, he "designed" a hybrid model: the animal Leyden jar.

Analogical hybrid models satisfy constraints of the target and the source domains [1]. In the animal Leyden jar case, the target domain constraints are satisfied as the biological object which fulfilled them is itself one of the components of the hybrid. The animal Leyden jar is a three dimensional object; its matter is a frog limb muscle in which the crural nerve is inserted (the animal model of the neuromuscular human system). On the other hand, constraints of the source model are satisfied in the animal Leyden jar, as the muscle is shaped like an electrical condenser (the figure of the Leyden jar).[6]

Constraints embodied in the animal Leyden jar were provided by the objects of the biological and physical domains that were part of Galvani's work environment. Following Donald Schön, it can be said that the physician researcher Luigi Galvani envisioned a coherence on the frog limb "land", land crossed by a nervous fluid circuit, in order to manage the problem relating to the "animal electricity"

[5] Experimental results relating to the fact that muscle contractions arises when a conductive arch is put over the prepared frog limbs, but nothing happens without the application of the conductive arch led Galvani to think that an *electricity inherent to the animal does exist.* "But when I brought the animal into a closed room, placed it on an iron plate, and began to press the hook which was fastened in the spinal cord against the plate, behold!, the same contractions and movements occurred as before. I immediately repeated the experiment in different pieces with different metals and at different hours of the day. The results were the same except that the contractions varied with the metals used (…). These results surprised us greatly and led us to suspect that the electricity was inherent in the animal itself" [11].

[6] Pietr van Musschenbroek from Holand, a famous experimentalist physicist of the Leiden University, was considered the creator of the first electrical condenser, the "Leyden Jar". Electricity stored within this jar is positive, and the one on the outer side is negative. They are separated from each other through an insulator (glass). When electricity is in balance there is as much positive as negative electricity at each side of the insulator. If electricity is unbalanced, the excess of electricity is repelled or relieved. If, for instance, water contained within the jar is electrified by a metal conductor introduced in it, unbalance happens and, therefore, an electrical discharge takes place. Electrical fluid can be put in circulation in such a way that equilibrium is restored, by a conductor arch linking the outer and the inner side of the jar [6].

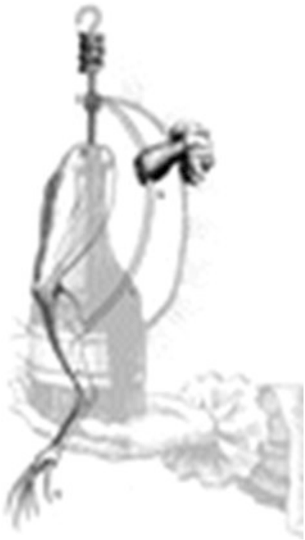


Fig. 1 Animal Leyden Jar (*Source* Piccolino [7], p. 345)

emplacement.[7] He used a physical device from his “*repertoire*” to do that: the Leyden jar [12]. It should not have been very hard for Galvani to figure the frog limb with the shape of a Leyden jar. As Marco Piccolino points out, the application of metallic wires hooks to the nerves or muscles of frog preparations (useful for administering electricity) contributes to envisioning specific parts of the animal organism as components of identifiable electric circuits [7].

2 The Cabinet: Factual Scenario of the Model Visual Structure

The “animal Leyden jar” model and a further adapted model, the “muscle fiber Leyden jar”, are physical objects that have a visual structure. On which is this one based? (Fig. 1).

Chandrasekharan and Nersessian highlight that the visual structure of an analogical model can come from an imaginary, counterfactual scenario: *“The visual structure of the model is not based on an existing structure recognized as the target; it is based on an imagined structure, a counterfactual scenario that is*

[7] According to Donald Schön, professional competence is a way of building the world that allows to manage problems, i.e., singular, uncertain and conflictive practical situations; and design is the main ontological process for the professional art exercise. In a general sense, design is the imposition of an own coherence. The designer, the prototype of which would be the architect, has a *repertoire* of examples to which he can resort to, in order to know and act. When professionals give sense to a situation taken as singular, they apply something of their *repertoire*.

implemented by the building process" [2]. I defend that (a) in some cases the visual structure of the analogy source comes from a factual scenario. In addition to this, I argue that (b) the latter does not imply that the visual structure is based on an existing structure recognized as the target.

(a) Galvani's case instantiates the idea that the visual structure of a hybrid source model can be based on a *real* environment. The constraints that satisfy the animal Leyden jar correspond to objects which *were* within the *cabinet* where he worked; furthermore, he combined and manipulated them jointly in many of his experiments. The visual structure of the analogical source emerged from these experimental arrangements (Fig. 2).

(b) 1. The object of the target domain modelled by the animal Leyden jar was the "prepared" frog. Although, in great measure, Galvani made experiments on it, the "prepared" frog did not determine by itself alone the conditions from which animal Leyden jar visual structure emerged. Furthermore, the possible pertinent structure of the "prepared" frog—the shape of the frog limb muscle, the way in which the crural nerve was inserted on it, and the site of the positive and negative electrical charges on the inner and outer surfaces of the muscle—came to be defined with the guide of the source hybrid model.

> It will not seem beside the point (…) that the muscle should be the proper seat of the electricity investigated by us, but that the nerve performs the function of a conductor [11].

So, in this case, the visual structure of an analogical source cannot derive from the pertinent visual structure of the target system.

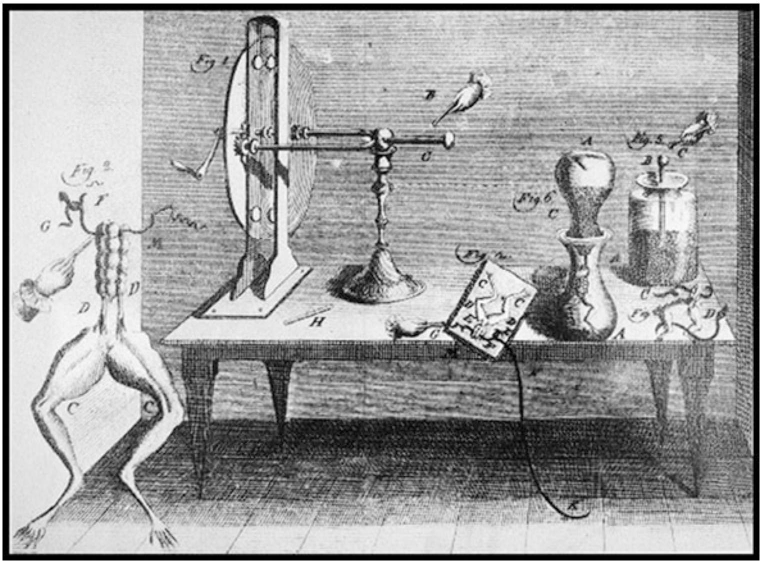


Fig. 2 Physical dispositives and "prepared" frogs (*Source* Galvani, [11], The first plate of the *De viribus electricitatis in motu musculari*)

(b) 2. Indeed, the pertinent space disposition of the target system, the "prepared" frog, was thought to be shown with the guide of an adapted source model. Although Galvani achieved the *conjecture* that the opposed components of electricity are in the muscle, giving arguments for his opinion; he was not able to find them *in the* muscle of the prepared animals limb. This left him unsatisfied. So he carried on a new strategy in order to "show" the place of the animal in which the opposed electricity is. He modified the analogical source: he designed a Leyden jar on a *muscle fiber* of the frog limb. This latter analogical source contained, as a component, an animal muscle fiber on which a nerve was inserted. The muscle fiber had a particular anatomical configuration: a substance, at an *isolating interface* between the inner and the outer surface of each muscle fiber. Galvani seemed to design the isolating substance as the glass that separates the inner and outer side of a Leyden jar. In this manner, the problem of representing animal electricity in muscle, but overwhelming the hallerian objection to the neuroelectric theory, was solved. The following step was to infer directly, from the visual structure of the constructed new hybrid model, the site in which the opposite electricity is placed [11]. This idea of inferring from physical models is developed by Ronald Giere in "Models as Parts of Distributed Cognitive Systems" [13].[8]

> But much less easily could he deny double electricity in one and the same muscular fibre who should see that it is neither difficult nor without some sort of truth that the same fibre should have external and internal surfaces opposite one another, either having observed a cavity, which some assign to it, or from diversity of substances, of which we have said it is composed, which cannot be without various holes and surfaces of the muscular substance [11].

Therefore, the pertinent space structure of the target system, the "prepared" frog, was shown with the guide of an adapted source model. The visual structure of this source did not derive from the pertinent visual structure of the target system, which, as the case study instantiates, can be unknown and problematic.

The "animal Leyden jar" was an analogical source with which Galvani seeked an explanation about the way of storing animal electricity in an animal and which physiological role this electricity plays was seeked. Knowledge obtained in this manner was meant to be transferred to the human domain, and, more specifically, to electric medicine. So, that model constituted a "nested analogy".

It was a hybrid model, that satisfied constraints provided by objects pertaining to Galvani's work environment. The fact that constraints embodied by the model corresponded to objects of his laboratory, and that the way in which those constraints were mixed in the model was suggested by arrangement of his work objects during experimental manipulation, allows to say that the visual structure of a hybrid source model can be based on a *factual scenario*. However, this does not imply that

[8] According to Giere's account, there are physical—three dimensional—models based reasonings, that can be understood as distributed cognitive processes instantiated by a system made of those physical models and persons. The mentioned models contain information and the humans extracts the seeked structure from them directly. It is not required to operate logically over "inner" representations, it is sufficient to interact in a physical and perceptual way with "outer" representations.

structure is based on an existing structure recognized as the target. The hybrid visual structure does not emerge from the bare conditions of the target domain object. It is rather the contrary what happens: the pertinent structure of the target domain object is defined and eventually exhibited with the guide of the source model.

References

1. Nersessian, N.: Creating Scientific Concepts. MIT Press, Cambridge (2008)
2. Chandrasekharan, S., Nersessian, N.: Hybrid analogies in conceptual innovation in science. Cogn. Syst. Res. **10**(3), 1–35 (2009)
3. Bertucci, P.: Therapeutic attractions: early applications of electricity to the art of healing. In: Whitaker, H.A., Smith, C.U.M., Finger, S. (eds.) Brain, Mind and Medicine: Essays in Eighteenth-Century Neuroscience, pp. 271–284. Springer, Boston (2007)
4. Cavazza, M.: Early work on electricity and medicine in the Bologna academy of sciences: Laura Bassi and Giuseppe Veratti. In: Pancaldi, G. (ed.) Bologna Studies in History of Science, 13. CIS, Bologna (2011)
5. Bresadola, M.: Medicine and science in the life of Luigi Galvani (1737–1798). Brain Res. Bull. **46**(5), 367–380 (1998)
6. Roller, D.Y., Roller, D.H.D.: The development of the concept of electric charge. In: Conant, J.B. (ed.) Harvard Case Histories in Experimental Science, Vol. 2. Harvard University Press, Cambridge (1964)
7. Piccolino, M.: Visual images in Luigi Galvani's path to animal electricity. J. Hist. Neurosci. **17**, 335–348 (2008)
8. Burggren, W.: Developmental physiology, animal models, and the August Krogh Principle. Zoology **102**, 148–156 (1999/2000)
9. Burggren, W., Warburton, S.: Amphibians as animal models for laboratory research in physiology. ILAR J. **48**(3), 260–269 (2007)
10. Focaccia, M., Simili, R.: Luigi Galvani, physician, surgeon, physicist: from animal electricity to electro-physiology. In: Whitaker, H.A., Smith, C.U.M., Finger, S. (eds.) Brain, Mind and Medicine: Essays in Eighteenth-Century Neuroscience, pp. 271–284. Springer, Boston (2007)
11. Galvani, L.: *De viribus electricitatis in motu muscolari. Commentarius.* De Bononiensi Scientiarium et Commentarii 7 (1791)
12. Schön, D.: Educating the Reflective Practitioner. Jossey-Bass, San Francisco (1987)
13. Giere, R.: Models as parts of distributed cognitive. In: Magnani, L., Nersessian, N. (eds.) Model-Based-Reasoning. Science, Technology, Values. Kluwer, New York (2002)

From Informal Thought Experiments to Agent-Based Models A Progressive Account of Modeling in the Social Sciences

Ruggero Rangoni

Abstract Thought experiments are widely employed in the social sciences, as many experiments are not affordable or even impossible to execute. Informal thought experiments, which are typical of classical economics, involve vagueness in the inference from their premises. On the other hand, mathematical models lack realism in their assumptions. Agent-based models are a particular kind of thought experiments, which are especially useful when our intuition is dealing with complex phenomena. We will argue that, contrary to mathematical models, agent-based models allow more realistic assumptions. On the other hand, unlike informal thought experiments, the conclusion of a simulation is the result of a strictly deductive procedure. Hence, agent-based models improve the realism of the assumptions of mathematical models, while avoiding the vagueness of informal thought experiments.

1 Introduction

Agent-based models are often presented as a new tool for the social sciences [3]. This approach leaves little room to compare them with more traditional analysis. One has the feeling that new aims and methods are at stake. In this paper, we will sketch a framework that allows us to consider agent-based modeling as a more conservative tool than it would appear at first sight.

In the second section, we will briefly introduce the method of thought experiments in the social sciences. Mathematical models will be discussed in the following section. In the forth section, we will take into account agent-based models. Finally, in the fifth section, we will compare the features of agent-based model with those of informal thought experiments and of mathematical models.

R. Rangoni (✉)
University of Rome La Sapienza, Rome, Italy
e-mail: ruggero.rangoni@gmail.com

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_26, © Springer-Verlag Berlin Heidelberg 2014

2 Informal Thought Experiments

According to Sorensen [15], a thought experiment is “an experiment that purports to achieve its aim without the benefit of execution”. Thought experiments are of common practice in natural and social science, as well as in philosophy of science, moral philosophy and even theology.[1] Nevertheless, there is little agreement on what kind of knowledge we can obtain through thought experiments, the debate being vivid since the traditional clash between rationalists and empiricists.[2]

When we run an ordinary experiment we gain new knowledge from observation. On the other hand, running a thought experiment simply means to imagine the consequences that would follow from a set of assumptions and some initial starting conditions. Apparently, from “armchair inquiry” we can at best infer useless tautologies and gain no new knowledge. However, we are rarely able to see all the logical consequences of a given hypothesis, or of a set of beliefs. This is why the conclusion of a thought experiment is often surprising. Hence, we might regard thought experiments as heuristic tools that help us gain insights from what we already know. In other words, as Kuhn stated, “[...] thought experiments give the scientists access to information which is simultaneously at hand and yet inaccessible to him” [9].

Thought experiments are run when the corresponding actual experiment is (i) not necessary (or at least believed so) (ii) unaffordable, for monetary or moral reasons, or impossible to execute. The former case is usually met when our aim is to show some logical flaw of an hypothesis. For example, think of Galileo’s well-known falling bodies thought experiment. According to the Aristotelian account, the speed of a falling body is proportional to its weight. If we imagine to tie together two bodies of different weight, how would their speed be affected? The lighter body will fall slower than the heavier one. Hence, the two bodies tied together will have to fall slower than the heavier one alone, as the lighter will slow down the heavier. On the other hand, the two bodies tied together will weight more than the heavier body alone. Hence, they will fall faster than it alone. The conclusion of this thought experiment is that two bodies of different weight, if tied together, would have to fall both faster and slower than the heavier body alone. As this is clearly absurd, the Aristotelian account is refuted. The actual experiment would be easy to carried out, but there is no need to do it.

Unaffordable or impossible experiments are common in the social sciences. In fact, as Mill [13] states with great clarity, society is far from being a neat laboratory, where all variables can be kept fixed but one, in order to observe its effects. Typically, a multitude of causal factors are present, over which we have little or no control. Hence, experiments can seldom be executed. To explain social

[1] For a taxonomy of thought experiments, see, for example, Brown and Fehige [2].

[2] Interestingly enough, despite their gnoseological commitments, empiricists like Locke and Hume made large use of thought experiments. Think, for example, of Hume’s [6] “missing shade of blue”.

phenomena, economists often have to rely on thought experiments—they isolate the causal factors in their mind, and imagine the possible consequences.

The method of abstraction and thought experiment has a long history in economics, and dates back at least to the eighteenth century. An easy and well-known example is how Adam Smith explains the development of the division of labor, imagining how it would work in a small tribe.

> In a tribe of hunters or shepherds a particular person makes bows and arrows, for example, with more readiness and dexterity than any other. He frequently exchanges them for cattle or for venison with his companions; and he finds a last that he can in this manner get more cattle and venison, that if he himself went to the field to catch them. From a regard of his own interest , therefore, the making of bows and arrows grows to be his chief business, and he becomes a sort of armorer. [] In the same manner a third becomes a smith or a brazier, a forth a tanner or dresser of hides or skins, the principal part of the clothing of savages. And thus the certainty of being able to exchange all that surplus part of the produce of his labour, which is over and above his own consumption, for such parts of the produce of other man's labour as he may have occasion for, encourages every man to apply himself to a particular occupation, and to cultivate and bring to perfection whatever talent or genius he may posses for that particular species of business [14].

Smith's hypothesis is that the division of labour might have emerged spontaneously. He proposes an easy thought experiment, which shows how it is in the interest of every individual to gradually specialize and pick a profession. Smith is vague about the details of the experiment, such as the size of the village or how much trading needs to be developed to get the process started. For the rest of this paper, we will address to such experiments as *informal thought experiments*, in which formalization and mathematics play a minor role—or no role at all.

The persuasive power of informal thought experiments is limited by their vagueness. In other words, we might be skeptical about the inference from the premises of a thought experiment to its conclusions. For example, think of Menger's account on the origins of money. Menger claims that money has spontaneously emerged from barter. His explanation proceeds as follows. First, different goods are salable in different degrees. According to Menger, we can say that goods are more or less salable, according to the greater or less facility with which they can be sold. Money is just a special case, being salable at the maximum degree. The subsequent step is to notice that it is reasonable to suppose that

> [...] when any one has brought goods not highly salable to the market, the idea uppermost in his mind is to exchange them, not only for such as he happens to be in need of, but, if this cannot be effected directly, for other goods also, which, while he did not want them himself, were nevertheless more salable than his own. By so doing he certainly does not attain at once the final object of his trafficking, to wit, the acquisition of goods needful to himself. Yet he draws nearer to that object. By the devious way of a mediate exchange, he gains the prospect of accomplishing his purpose more surely and economically than if he had confined himself to direct exchange [12].

Individuals who employ this strategy contribute involuntarily to the emergence of money, while they are only pursuing their own interest. In fact, Menger understands brilliantly that this simple strategy is sufficient to trigger a complex chain reaction. First, individuals employing this strategy are on average more successful than those seeking a direct exchange. Hence, the strategy will be imitated by others and become common practice. Moreover, as more individuals perceive a good as highly salable, the more they will be ready to accept it. This will contribute to make the good even more salable than before. Hence, once started, the mechanism is self-enforcing. At some point, one or few goods become universally accepted and maximally salable, i.e. they become money.

Menger's argument is persuasive, and yet not definitive. In fact, different hypothesis on how money might have emerged survive. Surely, it is possible to accept the reasonable premises of Menger's thought experiment—that some goods are more salable than others, that people are ready to imitate successful strategies, and so on—but to reject his conclusions. For example, we might claim that the chain reaction will never take place, or that it will only take place in very particular circumstances. Different people might have different opinions on the salability of goods. Hence, they might fail to coordinate even in the long run, and money never emerge. Will money emerge more easily in markets with many different goods or with a few of them? How large does the initial asymmetry among good salability need to be for the process to take place? Questions of this kind can't be answered if the experiment we propose is purely informal. At the same time, it would appear ridiculous to try to run definite thought experiments when we are dealing with complex systems. For example, Menger and Smith could have asked us to imagine one hundred individuals, each of them with a different attitude and behavior, and interacting with each other. However, any attempt to derive precise consequences from such assumptions is beyond our ability to compute huge amounts of data.

3 Mathematical Models

The scenario is very different with mathematical models, which have become the principal tool of mainstream economics since the first decades of the last century. In mathematical models theorems follow from a set of assumptions. Hence, unlike informal thought experiments, the inference from the premises of the model to its conclusions can't be questioned. However, mathematical models are not without troubles as they are often charged of being not realistic in their assumptions. In fact, every abstract model involves a number of assumptions that, strictly speaking, are false. Think, for example, of perfectly informed, olympic rational consumers and of profit maximizing firms. For the purpose of this paper, we will be content to characterize realism in the assumptions as what makes sense of the inductive leap from the model to the world. In fact, unrealistic models are often charged of being just mathematical games, which do not help us to understand the social world.

One possible line of defense, most notably pursued by Friedman [5] is to deny that models aim at explaining social facts. In his view, we should only judge models according to their predictive power, regardless of their truth value. However, anti-realism and instrumentalism in the social sciences are hardly defendable. At best, the empirical accuracy of abstract models is not testable. Moreover, one has the feeling that theoretical economists are really trying to grasp the mechanisms of the social world.[3]

If we want to give up anti-realism, we need a better account of falsehoods in models. Of course, we're not looking for models that are accurate in all respects to the target phenomenon. Models, and thought experiments more in general, are useful just because they differ from reality in the aspects that prevent us to manipulate it and run experiments. Think, for example, of the method of isolation. According to Mäki, while we run thought experiments we isolate causal factors in the same way we would physically do in a laboratory. In this sense, abstract models are unrealistic, but in a purposeful way [10].

However, in order to be mathematically tractable, models also distort relevant aspects of the target process or phenomenon. For example, think of convex utility functions, representative individuals and imaginary auctioneers. It's this kind of falsehoods that advocates of realism complain about. As distortions and idealizations come in degree, we would feel more confident in the conclusions of models if we could at least relax some of them. We have seen that both informal thought experiments and mathematical models have their drawbacks. On the one side, informal thought experiments are vague. On the other side, unrealistic assumptions typical of mathematical models make them questionable. Hence, there seems to be a trade-off in a model between being binding and being realistic.

4 Agent-Based Models

Agent-based models, also known as bottom-up models or artificial societies, are a special kind of computer simulations that study complex phenomena, thus being particularly useful in the social sciences. One defines the behavior of the agents and the features of the environment, in order to observe macro properties that eventually emerge, e.g. behavior patterns and equilibria in the population.[4] Following Epstein [3, 4], we can state that the main features of agent-based models are

1. The possibility to work with heterogeneous agents, rather than representative ones
2. The autonomy of the agents, as there is no top-down control over the simulation once it has started
3. The explicit representation of space, which makes agents interact locally
4. The possibility—and often the necessity—to work with bounded rational agents, rather than rational ones
5. The possibility to follow the dynamics of the simulation step by step, rather than focusing only on the final equilibrium.

[3] For a debate on Friedman's account, see Mäki [11].

[4] For an introduction to agent-base models, see, for example, Tesfatsion [16].

Table 1 Virtues and vices of different approaches

	Informal thought experiments	Mathematical models	Agent-based models
Assumptions	Realistic	Unrealistic	Realistic
Inference	Vague	Stricly deductive	Stricly deductive

Agent-based models are often presented as a new tool, with its own method and even with different aims if compared with more traditional analysis.[5] We want to argue quite the opposite—that agent-based models are a better way to do informal thought experiments.

Given some adequacy conditions, it has been argued that agent-based models are rational reconstructions of the processes that could have led to complex social facts, such as norms and institutions.[6] In this perspective, we could say with Robert Axelrod [1] that "agent-based modeling is a way of doing thought experiments in which the intuition of the scientist is aided by the computer."

Thought experiments run in our mind and proceed deductively. Does this hold for agent based models? It should be clear that being run on a computer is not essential of agent-based models, nor of any other computer simulation. This is true in the same way the checker-board is not essential to the game of chess, which is defined only by its abstract rules. The checker-board helps the players to remember the position of the pieces and to analyse the game. However, chess masters are able to play blind-folded chess. In the same way, we could imagine someone being able to run a blind-folded simulation, without the support of the computer.

Moreover, agent-based models are strictly deductive, each step of a simulation necessarily following from the antecedent. It is true that most simulations involve the use of random numbers, but we should remember that they are generated deterministically by the computer. A different kind of objection regards the fact that we often need to run a set of simulations, rather than one. In fact, it is common practice to vary the parameters and other initial conditions in order to test the robustness of our conclusions. However, it should be clear that the following statistical analysis is not part of the model. Rather, it is the first step of the inductive path that lead us from the model to the world.

5 Conclusions

The same problem in social sciences can be tackled with all three methods we've been discussing—an informal thought experiment, a mathematical model and a computer simulation. For example, let's go back to the origins of money. We have introduced Menger's account, but mathematical models and agent-based models

[5] For example, this is the position advocated by Epstein [3].

[6] A classic study of such adequacy conditions is Ullman-Margalit [17].

can be found in the literature as well.[7] This being true, we have grounds to compare virtues and vices of the three approaches. We wish to argue that agent-based models, if properly constructed, can be better-off than their correspondent mathematical models and informal thought experiments. In fact, they seem to allow for both

1. Greater realism in the assumptions, if compared with mathematical models. We have seen that agent-based models allow for more human-like actors, with limited rationality and information, heterogeneous among each other. We are also able to grasp the process, instead of focusing on equilibria, which are seldom met in the world. In our example, this would mean to allow for more human like agents and a sound process of exchange.
2. Strict deduction in the inferential process from those assumptions, eliminating the vagueness of informal thought experiments. This means that, unlike in informal thought experiments, if one accepts the assumptions of the model, is bound to accept the conclusions that follow. In our example, this would mean more transparency in the initial assumptions, as well as producing a binding argument.

References

1. Axelrod, R.: The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton University Press, Princeton (1997)
2. Brown, J.R., Feighe, Y.: Thought experiments. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2011)
3. Epstein, J.M.: Agent-based computational models and generative social science. In: Epstein, J.M. (ed.) Generative Social Sciences: Studies in Agent-Based Computational Modeling. Princeton University Press, Princeton (2006)
4. Epstein, J.M.: Remarks on the foundations of agent-based generative social-science. In: Tesfatsion, L., Judd, K.L. (eds.) Handbook of Computational Economics, vol. 2, pp. 1585–1603. North-Holland, Amsterdam (2006)
5. Friedman, M.: The Methodology of Positive Economics Essays in Positive Economics. Chicago University Press, Chicago (1953)
6. Hume, D.: A Treatise on Human Nature (1740)
7. Kiyotaki, N., Wright, R.: On money as a medium of exchange. J. Polit. Econ. **97**, 926–954 (1989)
8. Kobayashi, M., et al.: Simulation modeling of emergence-of-money phenomenon by doubly structural network. In: Nakamatsu, K. et al. (eds.) New Advances in Intelligent Decision Technologies, vol. 199, pp. 585–594. Springer, Berlin (2009)
9. Kuhn, T.S.: The Essential Tension. The University of Chicago Press, Chicago (1977)
10. Mäki, U.: On the Method of Isolation in Economics. Poznan Studies in the Philosophy of Science and the Humanities, vol. 26, pp. 316–351 (1992)
11. Mäki, U.: The Methodology of Positive Economics: Reflections on the Milton Friedman Legacy. Cambridge University Press, Cambridge (2009)

[7] For a mathematical model of the origins of money, see, for example, Kiyotaki [7]. Examples of agent-based models of the account are Kobayashi [8] and Yasutomi [18].

12. Menger, C.: The Origins of Money (1892)
13. Mill, J.S.: On the Definition of Political Economy and on the Method of Investigation Proper to it. Essays on Some Unsettled Questions of Political Economy. Batoche Books, Kitchener (2000)
14. Smith, A.: An Inquiry into the Nature and Causes of the Wealth of Nations. Oxford University Press, Oxford (1976)
15. Sorensen, R.A.: Thought Experiments. Cambridge University Press, New York (1992)
16. Tesfatsion, L.: Agent-based computational economics: a constructive approach to economic theory. In: Tesfatsion, L., Judd, K.L. (eds.) Handbook of Computational Economics, vol. 2, pp. 831–880. North-Holland, Amsterdam (2006)
17. Ullman-Margalit, E.: The Emergence of Norms. Oxford University Press, Oxford (1977)
18. Yasutomi, A.: The emergence and collapse of money. Phys. D **82**(1–2), 180–194 (1995)

Generative and Demonstrative Experiments

Tommaso Bertolotti

Abstract Current scientific practice is often identified with the experimental framework. Yet, what "experimenting" means could be less than perfectly clear. Going beyond the common sense conception of experiment, two broad categories of experiments can be tentatively identified: the generative experiment and the demonstrative experiment. While the former aims at generating new knowledge, new corroborations of hypotheses etc., the latter—which is actually the kind of experiment most laypeople came to terms with in their lives—is designed so that, by being successful, it reverberates knowledge on the experimenters/witnesses, thus instructing them, albeit the experimental outcome was well known beforehand. *Prima facie* the uninformed observer may not always be able to tell whether an experiment is generative or demonstrative, therefore the existing distinction must rely on something else, namely the framework they are embedded into. The concept of epistemic warfare, recently introduced by Magnani, can be of help in investigating this distinction, also to the scope of showing that it is not a sterile dichotomy but rather a theoretically fruitful continuum, and can help the analysis of epistemically relevant issues such as the repetition/replication of experiments and their potential failure.

1 Introducing the Experiment

The idea of experiment is intuitively connected with the common conception of modern science. Yet, until the second half of the twentieth century, philosophy of science reenacted the ancient bias against craftsmanship and focused chiefly on the theoretical aspects of scientific endeavor. Breaking this tendency, philosophical

T. Bertolotti (✉)
Department of Humanities—Philosophy Section, and Computational Philosophy Laboratory, University of Pavia, Pavia, Italy
e-mail: bertolotti@unipv.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_27, © Springer-Verlag Berlin Heidelberg 2014

milestones such as Hacking [10] and Gooding [9] claimed the experimental question rightly back in the epistemological feud, but the topic was quickly seized by a different branch of studies, sometimes called social epistemology, sociology or anthropology of science, which focused more on the social dimension of experimental settings (consider [3, 14]).

The aim of this paper is to make the best of these two approaches (the epistemological care for experimentation, and the social-anthropological outlook), in order to set the framework for an updated and consistent consideration of experiment: that is, what I mean to provide is an analysis of experimentation able to comprehend both *crucial* experiments carried out in laboratories, and the more modest kind of experimentation we came across, for instance, during our high school years.

I believe that experimentation is a particularly pivotal topic for the understanding of science as a whole. Science is a many-headed deity: Hacking [10] claims in the Introduction that, albeit one decides to deal either with scientific rationality or scientific realism, one topic necessarily ends up defining the other. There are many roads leading to the same castle.

1.1 Defining the Experiment

I shall begin by sketching out an extensive definition of what can be considered an experiment. In order to be fruitful, the definition of what an experiment is must be neither too broad nor too narrow: I would rather enumerate a list of features that, in my opinion, make up the experiment rather than providing a full definition.

- *Any experiment is characterized by a manipulative dimension.*

Hacking [10] stressed the importance of *intervention*, and rooted his scientific realism not on our possibility to *know* but on our possibility to *intervene*. This intervention has to be understood in its fully dynamic display: an experiment is not the result of the experiment itself, but the whole process by which this result is achieved—or not. This can be said of thought experiments as well: as contended by Gendler [8], a thought experiment like the ones conducted by Galileo cannot be reduced to a more or less sound logical argument, because the manipulation and hence the manipulability by an agent are the pivotal feature.[1] This is all the truer if we think of real experiments: you can *tell* somebody how an experiment was conducted (the preparation, the procedure, and the outcome), but the narrative and communicative reconstruction of the experiment, and the description of the involved procedure, are not the experiment itself. Experiments are a manifest example of *manipulative abduction* [18].

[1] This view on thought experiments is not universal. Some scholars contend that they can in fact be reduced to straightforward arguments [28].

- *All experiments have a transformative nature.*

Experiments can be thought of as mechanical systems: to begin with, we have what is being experimented on; we then add what the experimenters bring to the experiment, that is models, heuristics, techniques, personal hunches and so on. Finally, manipulation, intervention, the *work* can take place. This work has an output, of any kind. It seems to me that without this *output* there would be no experiment to begin with: the result can be new knowledge, so to say "extracted" with the experimental manipulation. Of course one cannot know that she extracted all the available knowledge from the experiment: just like when you squeeze an orange to have the juice, someone might show you there was more juice left, or that there can be a better way to squeeze it. Yet, as I will explain along the next sections, an experiment does not only generate (absolutely) new knowledge: the output of the experimental transformation might also be a new affordance (of the experimented, for instance), or new commitments (e.g., toward the advance of science). A High Energy Particle Collider experiment will try to produce new knowledge about subatomic particles, whereas a high school physics experiment might help students acquire a new and better understanding of a certain phenomenon, or a renewed commitment towards scientific progress: that, too, is an effect brought about by the kind of transformations enacted by the experiment.

- *All experiments are "situated".*

Any experiment begs for a situation. It could be argued that experimentation itself projects its situation: experiments are often about the controlled manipulation of a number of variables in order to see "what happens when...", or "if it is the case that *x*...". The laboratory is the situation *par excellence*, and the next subsection will focus on it, but it is not the only one: even in a non-scientific setting, but when people wish to make use of a kind of rationality that can be called scientific, the suggestion "Let's make an experiment" entails the setting of a boundary: it is pragmatic (i.e. deciding what matters and what does not matter), but also regards the assigned social roles. Depending on the interactions, the peers of the person who calls for an experiment will be involved as active participants (as in, "Let's make an experiment: what would you do if...?"), or as onlookers/witnesses, expected to provide some opinion about the conclusion and the procedure. In any case, the experiment takes place within a well defined place, which can be more or less physically determined. Hacking refers to "mature" laboratory sciences those "in which we investigate nature by the use of apparatus in controlled environments, and are able to create phenomena that never or at best seldom occur in a pure state before people have physically excluded all 'irrelevant' factors" [11, p. 507]. Let us therefore take a closer look at what we called the boundaries of the experiment, which must be the boundaries of the laboratory.

1.2 Setting the Boundaries of the Laboratory

Let us accept our loose intuition about what a laboratory is: informally, we can think that a laboratory is the specific location where scientific experiments take place. But more can be said about the lab: first of all, what are the actual boundaries of the laboratory, understood as the *lieu* of the experiment? With this respect, I invite the reader to make a small recollection and consider what the word "laboratory" makes her think of: she could think of the instrumentation used for experiments, she could think about the instrumentation strategically laid out on workbenches, and perhaps of scientists carrying out experiments on these workbenches. Plus, thinking of the labs she might have attended, she could also think of all this *and* onlookers standing by and witnessing the experiment.

It is clear that the epistemological consequences of where we set the boundaries of the lab are quite significative. In the last item, the onlookers can be students, colleagues, sponsors, visitors at a science exhibition, and so on: I do not believe that including them among the possible target of the knowledge *transformation* enacted by the experiment necessarily means to shift the investigation from the epistemological plane to a social, anthropological one. Science as an *actual* human endeavor cannot be investigated excluding the human dimension it relies on. Knorr Cetina wrote that "the power of laboratories (but also their limitations) resides precisely in this 'enculturation' of natural objects. Laboratory sciences subject natural conditions to a 'social overhaul' and derive epistemic effects from the new situation" [3, p. 28].

I believe that the soundness of Knorr Cetina's statement does not exclusively follow from the adoption of an ethnographic outlook on science.[2] Conversely, it is easy to understand that the *enculturation* and the *social overhaul* are mutually implicating in our conception of science, because laboratories are more than a set of instruments, and even more than a set of affordances displayed by those instruments: labs prevent scientists from having to study a natural object "*as it is, [...] where it is [and] when it happens*" (p. 27). In this sense, labs allow the manipulation of the object far better than the natural context would (if it would at all). Even scientific models, playing a pivotal role in the economy of the lab, partake of a similar nature, being on the one hand highly manipulative human construals (therefore cultural too), while on the other they are necessarily bound to the natural object: in fact models are fundamental in binding the modeled natural object into a specific phenomenon [1, 19].[3]

The laboratory includes the experimenters as well, inasmuch as they are not separated from what is being experimented. Albeit Hacking is thankful toward the "large number of studies by philosophers, historians, and ethnographers of

[2] Also Nersessian's outlook on science is often characterized by a particular attention—called "ethnographic"—to the actual dynamics at play in a laboratory (cf. for instance [25–27]).

[3] Hacking [10] contends as well that many phenomena come to happen uniquely as they are *created* in laboratories.

experimental science," [11, p. 508] he seems less eager to concede a more significative role to human intervention per se, which is conversely mentioned by Heelan by means of the "instruments, standard procedures, experimental skills, laboratory traditions, and the social context of the research community" [7, p. 525]. Scientists are not simply interchangeable operators: two teams working on the same raw objects would not apply the same methodologies or necessarily obtain the same results. Indeed, "not only objects but also scientists are malleable with respect to a spectrum of behavioral possibilities. In the laboratory, scientists are methods of inquiry; they are a part of a field's research strategy and a technical device in the production of knowledge" [3, p. 29]. It seems to me that what could be at stake here is not the dispossession of epistemology by social approaches, but rather the opposite, that is the epistemological flooding of some aspects of scientific endeavor which—by default of better option—have so far been labeled as social but do rather concern an agent-based and factual approach to science, also leaning on an *actually happens rule*.[4]

Why then should we set the limits of the laboratory at the experimenters' level, and not admit the onlookers as well? Why should the "social overhaul" advocated by Knorr Cetina only involve the experimenters? As I will show in the following section, some experiments do not benefit the experimenters at all, in a strict sense, inasmuch as they produce knowledge that had already been acquired, and yet the same experiments cause indisputable epistemic effects on those who *observe* them. For them the experiment still produced a valuable transformation of knowledge, by which they gained a new *understanding* 1) of the phenomenon that was *explained* to them through the experiment, but also 2) of how scientific rationality works. It could be contended, in fact, that some experiments carried out at schools or science exhibits serve the chief purpose of exemplifying some tenets of scientific method.[5]

Concluding this section, we might say that there are indeed many kinds of laboratory, in which different kinds of experiments take place: there are labs for basic research, industrial labs, labs for medical research, and then there are laboratories in schools, science museums, and also the laboratories shown in

[4] This rule was introduced by Gabbay and Woods as a tenet of their new approach to logic, referring to the fact that logic should model how real agents think: one should try to correct the model so it fits the facts, and not try to amend or obliterate facts to make them fit the model [6, 35]. In this context, I use it to suggest that philosophy of science should match what science really is, and not arbitrarily cut out aspects of the problem by labeling them as external to the analysis (for instance, "social").

[5] The onlooker's gain of a renewed commitment towards science, be it specific for a particular research/discipline or to scientific endeavor in general, is just as vital for the development of science as the generation of new knowledge through experimentation. Contemporary *knowledge societies* massively rely on the development of science, which in turn relies on the will of citizens to care and spend for it [17]: funds are just as vital as genius and intelligence for the survival of science. This view is coherent with Magnani's conception of science as an *epistemic warfare* [19], which also includes non-epistemic strategies that are nevertheless crucial for science, such as those for the dissemination of knowledge, the acquisition of funding and so on.

educational TV programs: most of laypeople are acquainted with the latter kind of laboratory, that is the physics/chemistry/biology lab at high school, or those they see in science museums or on the Discovery Channel. Such acquaintance fuels our *thinking to know* what every lab should be like, which is in fact a *hasty generalization*.

In his classic book *Science in Action*, Latour enacts his anthropological approach to science narrating the epistemological adventure of an anthropologist taking a full immersion in the scientific endeavor. Interestingly, he makes the narrator say: "We came to the laboratory in order to settle our doubts about the paper, but we have been led into a labyrinth" [14, p. 67]. Specifically, the doubts referred to a reading of endorphin levels, which had to be interpreted through graphs and indicators, yet this bewilderment is common to many onlookers approaching a scientific setting: *we came, we saw, and yet we have not understood anything*. And yet, we saw experiments, at school, at the science museum, on TV, how comes?

In sum, experiments take place in laboratories, and laboratories may include onlookers. Yet not all experiments are geared towards onlookers the same way: to certain experiments anyone can be an onlooker and benefit of the epistemic effects, to others the onlooker is defined by very specific characteristics. In my opinion, this depends mostly on the kind of experiment at stake. If different kinds of experiment exist, it is legitimate to wonder how many kinds there are, and how we can tell them apart.

2 How Many Kinds of Experiment Are There?

A kind of *taxonomy* of experiments is not unusual among philosophers of science, and such differentiations sometimes merge into other connected ones. Gooding, for instance, links the concept of experiment to its *reconstruction*, obtaining six different kinds of reconstruction to be employed in different narratives: namely *cognitive, demonstrative, methodological, rhetorical, didactic, and philosophical* [9, p. 7], each with their peculiar scope. Notwithstanding the utility and soundness of this differentiation, I contend that its root lies at a lower level, and actually underdetermines it. The whole spectrum of experimental activity, as far as natural and model-based sciences are concerned, could in fact be reduced to two major forms of experimentations. One of the advantages of this proposal, which I dub a "plea for epistemological austerity", is that every distinction causes some unhappy left-outs: Steinle [33], for instance, lamented that the "standard view" in the Nineties of the past century would disregard as epistemologically irrelevant those experiments that were aimed at discovery—and not at the test of a clear hypothesis, or at the retrieval of a particular measurement. Such conception would in fact leave out a number of fundamental instances in the history of science: grouping the experiments into two sets, namely "generative" and "demonstrative" experiments could instead cause a lesser number of *homeless* instances.

2.1 The Generative Experiment

I could begin by suggesting that what I call the "generative" experiment is the kind of experiment that common sense has acquired, but this would be misleading. I contend, indeed, that the common-sense conception of experiment is somewhat blurred, so that the generative experiment, which is what we *should* think of when we think of a scientific experiment, does not coincide totally with our intuitive conception of experiment.

The *generative* experiment is the experiment whose outcome is *not known* beforehand, and its aim is to manipulate and transform the *experimentandum* (what is being experimented on) into knowledge that is new for everyone. To put it another way, it is the kind of experiment where the epistemic target,[6] that is what the experimenters want to obtain, is intrinsic to the experiment (this latter claim might seem a truism, but the next subsection should prove the opposite).

Most experiments in the history of science can be thought of as *generative* experiments. It is the kind of experiment where you *test* something (a hypothesis, a theory), and is usually comprised within a theoretical framework. It is also true that "one can conduct an experiment simply out of curiosity to see what will happen" [10, p. 154]: not only experiments that are well nested within a particular theoretical framework, for instance those aimed at testing a particular hypothesis, or at finding out a particular measurement (think of Millikan's experiment, projected to measure the elementary electric charge), but also entirely "exploratory" experiments are generative. According to Steinle, explanatory experiments do not rely on a "specific and well-defined procedure, but [include] a whole bundle of different experimental strategies", and their "central epistemic goal is the search for general empirical rules and for appropriate representations by means of which they can be formulated" [33, p. S73].[7]

But also in our everyday life, when we make use of scientific-experimental rationality to put some makeshift model to the test, we recur to generative experiments to gain some new knowledge. I can send myself an email to see if my IMAP server is really experiencing issues, and I can ask a friend to email me as well. I can put a five-dollar bill in a vending machine to test it before butting a twenty-dollar bill, to see if the machinery works properly. Generative experiments are often conducted as part of model-based activities: I can ask a relative to simulate a social situation to gain better knowledge about some possible

[6] I specify epistemic target, as the scope of the experiment, to differentiate it from Hacking's use of the word *target*, by which he refers to a part of the *"materiel"* of the experiment (cf. [11, p. 509]).

[7] Steinle's aim in describing exploratory experimentation is to allow the appreciation of the epistemological importance of this kind of experiment, while the "standard view" tended to disregard them as part of epistemically irrelevant *discovery* processes. Exploratory experiments are particularly relevant for entering new fields requiring new concepts and new general facts [33]. The explanatory experimentation can also be extremely tacit, and consist chiefly of "thinking through doing" [16].

consequences of an action of mine, or a man might cast small objects off a table to assess the likelihood of himself surviving after jumping from a cliff with his car. In those cases, what I gain from the manipulative intervention of the target (that is from the experimentation) is some knowledge I did not possess before.

In sum, the focus in generative experiments should be put on their ability to intrinsically produce new knowledge. This is the kind of experimentation that engages theory (and theories): as suggested by Steinle [33] and Hacking [11] among many others, some experiments—which I label as generative—can precede theory inasmuch as they can illuminate new fields of scientific research and provide it with new concepts.[8]

With respect to this kind of experiment, even scientific common sense knows that theories should behave according to the already mentioned "actually happens rule": experimental observations affect theories. Experiments are where theories can be falsified [29], and experiments that do not go as expected can affect the scientific paradigm, taking it to an eventual crisis [12], or causing scientists to fix the protective belt of the program to keep it progressive [13]. In the next two subsections I will show how only one kind of experiments indeed affects theories, and then move to analyze a wide and yet peculiar class of experiments, that—even though they can be called experiments to their full right—are not expected to affect theories at all.

2.2 The Demonstrative Experiment

It is now time to deal with an apparent contradiction: we know that experiments are, so to say, the field artillery of scientific progress, and it is on experimental grounds that new knowledge is either discovered or validated. On the other hand, we also know very well that most experiments we—as laypeople—witnessed (even in decent laboratories) did not add anything to scientific knowledge. It would not be right to arbitrarily exclude them from the category of experiments, because they display all the traits I pointed out in Sect. 1.1, and they also fit with the more demanding description proposed by Hacking [11].

I am referring to most experiments carried out in schools, exhibitions, museums, and so on. For instance, they can be experiments aimed at demonstrating or

[8] Hacking suggests several examples from the actual history of natural science that refute Popper's claims according to which "theory dominates the experimental work from its initial planning up to the finishing touches in the laboratory" [10, p. 155]. The debate on the theory-ladenness of experimental facts is often brought to quasi-metaphysical issues: one way to tackle it is to appeal to the intuitive notion of theory (as folk theory). Experiments may precede particular theories, and yet rely on past sub-theories about substances, agency, causation etc. Thus, to say that an experiment precedes theory—and so does the experimental observation that follows such experiment—does not indeed equal saying that the experiment generates new coherent knowledge *ex nihilo*. After all, we could claim that intuitive, hard-wired theory precedes even out every-day observation, even at the lowest levels of the perception of images, sounds etc. [30].

illustrating a law or a theory, fostering a better understanding of it. With this respect, at least in the Italian school system, theory overwhelmingly precedes experimentation: in chemistry or physics courses, experiments are not even used to stimulate theorization upon the students' minds, but rather as a persuasive proof to show that what was explained in theory *actually happens*.[9]

This kind of experiment could be thought of as *deduced* from theory in a strong sense, opposed to the *weak* Popperian sense of experiments *informed* by theory: The procedure of the experiment is vouched for by the theory it means to put in display. Is it a *paetitio principii*? Not really. Consider this example:

1. *Experiment E* (for instance Maxwell's or Faraday's experiments on electromagnetism) is crucial for the establishment of a *Theory T*;
2. *Theory T* is established;
3. *Experiment E** is used at school to prove the adequacy[10] of *Theory T*.

*Experiment E** is a (usually easier) version of *E*, updated according to the theory it means to demonstrate. If its real aim was to *test* the theory, then of course it would be begging the question. But who would expect high-school students to be *actually testing* a theory? Everybody knows that high-school level optics, or electromagnetic physics and so on *do work*. Proving it *n* more times every day, in *n* school laboratories, does not add one bit to the robustness of those theories. *Experiment E** aims at providing students with an actual proof that what they studied (or they are going to study) is really so.

Even if you think about experiments that do not aim at demonstrating a law, but rather at isolating a phenomenon so that it can be shown for some theoretical scope, the defining element is that the experimental *outcome is known beforehand*.[11] Contrarily to the *generative* experiment, in this case the epistemic goal is extrinsic to the experiment itself: it means little to say that the experiment *in se* was successful, because it was planned to be successful. The experiment is successful in its *actual* scope if it operates any *change* within the epistemic configuration of the observer, after she witnessed the positive (staged) outcome of the experiment. That is to say, the experiment is successful if it triggered a new awareness in the observer, for instance a student might be further persuaded about the empirical adequacy of a theory, or a citizen might reconsider the importance of electing a prime minister advocating more funding for scientific research. Or, simply, their aim could be to convey indeed a bit of *local* knowledge about some

[9] This concept is well exemplified by a sign hung in my chemistry laboratory at high school, which would read something along the lines of "If I listen I will forget; if I see I will remember; if I do I will understand". The experimental dimension is taught as completely subsidiary to abstract theory.

[10] Please understand this word in an intuitive sense, as in "What they taught me about the *Theory T* does indeed happen in real life", and not as laden with implications about the epistemological debate about the truthfulness or acceptability of a scientific theory.

[11] This claim clearly begs for some considerations about the *failure* of an experiment: I will address this issue in Sect. 3.2.

phenomena, but on the overall, to infuse the belief that science is "interesting", or just "cool".

This class of experiment could be defined as *demonstrative* or *explanatory*, contrasting it with *generative* experiment. Interestingly, one could say that in their scope of disseminating scientific knowledge (for various purposes), demonstrative experiments have become more and more widespread together with the growing impact of science on society. Living science shows in the eighteenth and nineteenth century, analyzed by Raichvarg [31], provide a clear example of a *demonstrative* experimental framework, which could be seen as the ancestor of modern science exhibitions or scientific shows for general audience on TV. One of the scientists/showmen mentioned by Raichvarg would start his experiments with the following call:

> And if I am here among you, it is because all of you must draw from my demonstrations, the true and natural principles of the forces which are above us, these forces which frighten the ignorant but supply the educated with all the moral pleasures of intelligence (p. 3).

Raichvarg draws from his analysis a list of characteristics that were typically common to science shows, and still apply to scientific dissemination aimed at general public:

> – They reach a wide audience, an audience which could be defined as a public with no scientific training...They come to the fairground for anything but science, but then they meet science face to face...
>
> – The importance of the current events of science, mostly because on a fairground one must astonish everybody to attract everyone and get your pennies back...
>
> – A continuous desire for good pedagogy, together with a continuous desire for wonder, if not for the supernatural! (p. 4).

These *experimental shows*[12] did not contribute to form scientists, just as contemporary science classes at high school do not mean to train scientists, and science museums do not either [15]. On the other hand, these forms of dissemination do play a pivotal role in educating people that might undertake or value a scientific career. A living science show, just as a school experiment, may indeed induce in the observer a taste for scientific methodology, or just make her aware of its existence.

A final question concerning the demonstrative experiment might arise: since we are accustomed with experiments carried out at school, often with obsolete equipment, it seems that the difference between a generative and a demonstrative experiment should be most easily noticed. In my opinion, from a phenomeno-

[12] The expression is a bit of an oxymoron, but it means to stress the *staged* dimension of many demonstrative experiments. Concerns about the esthetic dimension of their replication will be addressed in Sect. 3.1.

logical point of view, it is not so. We should not be fooled by the *time lag*: if most of our school labs look like museums of past century science it is just because those instruments were once upon a time the cutting edge of generative experimentation (think of Volta's battery and most electromagnetism-relate devices). If we removed this time lag, which is merely contingent, we would be phenomenologically unable to tell one kind of experiment from another, if not by considering the cultural and social framework an experiment is nested in.

Let us make a quick thought experiment: imagine in the near future a highly-funded high school in some advanced country, whose politicians place a great emphasis on education. Just as our high schools have a physics lab, that high school as a High Energy Particle Collider in its basement, and teachers use it to instruct pupils about quantum physics. If the same-old-friendly alien landed on Earth and could not understand human language, and witnessed the experiments carried out in that school, and those carried out at CERN (for instance), it could not be able to tell any difference: what goes on, apparently, is the same. Yet, once our alien managed to set its intergalactic translator to understand our language, it would see at once the difference, since the HEPC at school would be embedded in a pedagogical framework of demonstrative experiments, whose outcome are already known by the teacher who can therefore lead the pupils along the right path. Time lag, and thus the obsolescence of experimental materials, are not a necessary criterion to tell a generative experiment from a demonstrative one, since *prima facie* they cannot be told one from another, unless considering—as I said—the setting they partake of.

3 Consequences of the Distinction

Now that the distinction between generative and demonstrative experiment is in place, I will use it to tackle two epistemological problems, namely the repetition/replication of experiments and their failure. Once again, I will try to match common-sense expectations with the actual scientific practice, past and current.

3.1 Differentiating Repetition and Replication of Experiments

In an interesting paper about the conception of experiment repetition in the past centuries, Shickore [32] sets out stating that "[t]oday it is generally assumed that isolated experimental outcomes—'one-offs'—are insignificant. Twentieth-century philosophers of science, most notably Popper, made the reproducibility of experimental results the basic methodological requirement for successful experimentation: if an experiment cannot be re-done, it is invalid" (p. 327–328). Indeed, the possibility of re-doing an experiment became one of the first tenets of contemporary scientific rationality.

Before applying my distinction (between kinds of experiments) to the problem of redoing experiments, a brief semantic interlude is required, which—I think—might let the reader foresee my claim before I make it clear. It is sometimes said that experiments are "repeated", while sometimes they are "replicated". I believe that the two terms can be sensibly separated, each with its own proper meaning.

- **Repeating** an experiment exemplifies the epistemological tenet towards the re-doing of experiments. You *repeat* an experiment when you put the known outcome between brackets and proceed entirely as if it was unknown. The focus of repetition is on *what* outcome will be obtained, and *whether*—changing certain factors—the same outcome will be obtained again.[13]
- **Replicating** an experiment focuses on the replication of the procedure and not only on obtaining the same outcome. You *replicate* an experiment without necessarily putting the outcome between brackets, because what matters is observing *why* a particular procedure yielded such an outcome. Once the reason is found out, it is possible to replicate the experiment with the pragmatic certainty that if the outcome differs from what expected, then a mistake was committed in the procedure.

I suppose that such semantic characterization foreshadowed quite clearly the rest of my argument. As for generative experiments, I think that *repetition* is the case. Repetition engages the intrinsic epistemic goal of the experiment. Repeating the experiment does not mean necessarily to redo the same experiment over and over. This is what happens every day in schools worldwide, and we know that it has little epistemic value for the progress of science.

> A number of scholars have stressed that scientists rarely try to copy the exact same experiment. Rather, experimenters seek to obtain similar results in different experimental settings, and experimental results are considered valid if multiple determinations of the evidence are possible. The crucial notion here is reproduction by *doing something different* [32, p. 328].

The repetition of an experiment in a generative epistemic context is valuable because it may challenge the previous outcomes of the same experiment, for instance it can interfere with claims of universality (by "doing something different"). Repetition has therefore chiefly epistemic concerns.

Repetition can indeed be about the *same* experiment, but in this case it is about looking for *freak factors* of the experimental procedure, and make sure that the result is accurate. Even if scientific truths are notably *provisionally* true, the search

[13] To make students assimilate this concept, physics teachers often deploy plethoric lists of settings (e.g. here, at the Equator, on mount Everest, on the Moon, on Mars, in a billion years, and so on) where a law (such as "All metal bars expand when heated") must apply for it to be universal. The different settings correspond to a series of real or potential repetitions of one or more experiment concerning the law in question.

for freak factors will end at some point. No branch of science still heats metal bars every day on normal conditions to see if they expand and by what coefficient.[14]

Whereas such use of repetition was already in vogue in Early Modern science, its role was chiefly to corroborate (and make appear *as reliable*) one's own experimental results by the method of the *slight modifications*: Schickore, building his case study on an Italian eighteenth-century microscopist and physiologist, states that "Fontana's methodological thought is particularly interesting because he stressed the importance of *repetition* of his own experiments. The text is packed with claims that experimental trials were repeated 'a hundred times' or even 'a thousand times,' and that thousands of animals were used. Also, the experiments were varied 'in a thousand ways"' [32, p. 328]. Only subsequently the stress was placed on the assessment (via repetition) of experimental results obtained by other scientists.[15]

If the repetition is meant to engage the outcome of some other scientist's experiment, then again it can partake of a generative nature. Assessing someone else's experiments is, as a matter of fact, one of the pillars of contemporary scientific practice: a purpose of publishing experimental procedures in peer-reviewed journals is to offer the experiment to the assessment of peers, so that other scientists can repeat it and see *if* they obtain the same results. With this respect, an experiment is scientific if it is available for repetition, so that somebody else can repeat it and—perhaps—falsify its previous outcome: it is not necessary, for an experiment to be deemed *scientific*, to obtain necessarily the same outcome upon every different repetition.[16] Also thought experiments, inasmuch their repetition does not lead *necessarily* to one indisputable result, can be seen as generative in their repetition [2].

As far as demonstrative, or explanatory experiments are concerned, it follows from the initial argument that we should be mostly dealing with *replication*, for a number of reasons. First of all, whereas the redoing of generative experiments has epistemic concerns (since the previous outcome is what has to be challenged), the redoing of demonstrative experiment must face different constrains: indeed, their outcome is already known and their scope is to disseminate knowledge for the

[14] "Scientists do not repeat the same experiment *ad nauseam*. They perform an experiment a 'sufficient' number of times (whatever that might be), and then perform it no more. The experiment becomes a part of history, to be performed again, if at all, only by science students as an exercise" [23, p. 248].

[15] This conception was rather absent in early modernity: "Recent methodological frameworks highlight robustness, the importance of multiple determinations of experimental outcomes through a variety of independent procedures. While some parts of Fontana's project could perhaps be reconstructed in hindsight as multiple determinations of experimental results, neither he not Redi [a physician and naturalist at the court of the Grand Duke of Tuscany] explicitly called for independent determinations by different means to make an experimental result more reliable" [32, p. 344].

[16] Of course, in the latter case, something must be wrong either in one of the procedures, or in the theorization on which the experiment relied. About this issue, see Sect. 3.2.

benefit of the observers, therefore their peculiar constraints are chiefly *esthetically oriented*.

I do not mean this in a strong sense, *à la* Feyerabend: it is not that experiments carried out in contexts of dissemination are a work of rhetorics. My contention is that the will to reproduce a successful experiment may focus the attention on the reproduction of the *same* procedure, which therefore acquires a ritualized dimension that laminates the epistemic concern.

As a matter of fact, being certain about the outcome (be it an experimental result or an experimentally-confirmed theory) causes a shift in the perspective: the objective is not to redo the experiment to see what happens anymore, but to replicate it in the most convincing and understandable way. This can also be said of actual scientific experiments: sometimes, in the reconstruction of a discovery, when things seem to go too smoothly, it may be the case that a more pleasing *demonstrative* experiment was smuggled in place of the original *generative* one.

Interestingly, Hacking reports an annotation of Maxwell's about the work of Ampère which sums up quite neatly the essence of the replicated demonstrative experiment:

> We can scarcely believe that Ampère really discovered the law of action by means of the experiments which he describes. We are led to suspect what, indeed, he tells us himself that he discovered the law by some process he has not shewn us, and *when he had afterwards built up a perfect demonstration he removed all traces of the scaffolding by which he had raised it* [10, p. 162, added emphasis].

This methodological reconstruction is akin to the one I put forward in Sect. 2.2, by which the demonstrative experiment is somehow deduced from a theory already confirmed as adequate. Consequently, this kind of experimentation (already drawn from a successful experimental confirmation) is ready for replication without excessive worries about the outcome, but rather about its development: if the experiment is carried out correctly, it will be successful and prove our initial hypothesis.[17]

[17] Furthermore, Schickore seems to connect the early-modern care for repetition *in se* with a chiefly demonstrative dimension: "References to multiple repetitions have been interpreted as an echo of an Aristotelian conception of experience; as a literary device to bolster an experimental report; as a literary tool to highlight the wealth of the experimenters' patrons; or as an expression of a general commitment to experience that marked the beginning of modern experimental science" [32, p. 329]. Such an understanding of repetition clearly embeds it in a demonstrative framework akin to the non-epistemic strategies advocated by Magnani's *epistemic warfare* (see footnote 5). Schickore also hints at how repetition, in Galileo, served as a conceptual wrapper to *run* experimental observations as general facts: "Claiming results that accrued from trials repeated 'a full hundred times' was a way of saying 'things *always* behave this way,' and hoping that the reader would believe it" [4, p. 134].

3.2 The Meaning of "Failure"

Repetition and failure are strictly interconnected. As I suggested in the previous subsection, repetitions of *generative* experiments are aimed at testing the outcome of the experiment (and so at testing the hypothesis, theory or measurement that had been carried out during the experiment):

> Our ability to recognize when data fail to match anticipations is what affords us the opportunity to systematically improve our orientation in direct response to such disharmony. Failing to falsify hypotheses, while rarely allowing their acceptance as true, warrants the exclusion of various discrepancies, errors, or rivals, provided the test had a high probability of uncovering such flaws, if they were present. In those cases, we may infer that the discrepancies, rivals, or errors are ruled out with severity [21, p. 18].

If "[a] test 'uncovers' or signals the falsity of *H* by producing outcomes that are discordant with or that fail to 'fit' what is expected were *H* correct" [22, p. 352], then it sparks a procedural loop involving a careful check of the experimental conditions (looking for freak factors), a revision of the hypothesis or—ultimately—a revision of the model [1, 19]. Therefore, in case of experimental failure, the existing tension between the experiment*er* (and her background knowledge) and the experiment*ed* is resolved in favor of the latter, and thus the dignity of the falsifying failure is *respected*. Failure becomes yet another manipulative factor at play in a subsequent experiment. Failures are able to climb back over the experimental framework and crawl inside of general theories from one minimal experimental discrepancy.

> When we falsify a prediction, however "local" it is, we falsify whatever entails that prediction, however general or large-scale. There is, in this respect, no localization of the refuting process. The fact that we may try to find out which part of the refuted whole is to blame is another question–the Duhem question [24, p. 105].

Think of how the inaccurate predictions fostered by Newtonian mechanics about the orbit of Uranus jeopardized the adequacy of Newton's theory in toto: this failure was accepted by Le Verrier, and transformed into new knowledge that managed not only to preserve the adequacy of the theory but also discover a new planet, Neptune.

I suggest, though, that in particular (yet scientific) settings, namely in demonstrative experimental frameworks, Musgrave's claim [24] is wishful, or at best it is the object of a mere lip-service. That is, sometimes a "local" falsification does not affect what entailed the falsified prediction at all. Experiments carried out in schools, for instance, can "not work out" for a number of reasons, in a more or less meaningful way (the phenomenon may not occur, or some measurements might be different). What happens in this cases? Nothing at all.

When a demonstrative experiment fails, the general/expected outcome of the experiment is not questioned. This peculiar "experiment *token*" may have failed, but not the "experiment *type*" it stood for [23, p. 252]. Failure is made into something relative to this peculiar occurrence: it is a matter of *here an now*—this particular experiment failed, but by no means it falsified the theory it was meant to prove. This can be supported by a dialectical interplay with the observers, aimed at illuminating and then filling ignorance bubbles with demonstrative *emergency knowledge*: this process is usually introduced by rhetorical questions along the lines of "Okay, you know why the experiment didn't work out?", followed by information—often in-between *ad verecundiam* arguments and plain magical thinking—about the involved instrumentation, secondary phenomena affecting the materiel involved and so on. I label this filling as *magical* because the leading experimenter is saving the expectations of the others by strategically deploying information that was only in her background knowledge: sub-experimenters (for instance pupils, or laypeople visiting a science exhibition) lack the necessary background to make sense of this information, which is therefore offered as self-justifying, or rather justified by the authority of the leading experimenter. There is a significant *appeal to authority* at work in the dissemination of scientific knowledge, even if the latter is presented as immune to authority constraints. Furthermore, it could be said that this *authority overhaul* is *necessary* if only to convey and evoke commitment towards scientific method and its *unconstrained nature*.

One last epistemological effect of this mechanism is worth noting: constructing his argument against the fictionality of models, Magnani [19] contrasts a static understanding of science—for instance the one conveyed by textbooks—with the actual understanding of the dynamic nature of scientific endeavor, and states that if they are seen statically then of course models appear as fiction. The demonstrative experimental framework I described raises the stakes. Demonstrative experiments seem to entail the kind of *fictionalism* that sees models as fictions depicting missing systems [20, 34]. Why? Consider failure in a demonstrative experiment: the unexpected wrong outcome is injected with emergency knowledge ("I am telling you why the experiment did not work out"), and so the model indeed appears as an awkward fiction (the phenomenon that the model should actualize does not happen). Furthermore, a demonstrative failure turns the observed reality into a fiction as well (a missing system, "which you should have seen in the experiment but you didn't..."), in order to support the cost of the what was to be demonstrated (be it a model, a law, etc.). In case of failure, the tension is resolved in favor of the experimenter and her background knowledge.

What is the final result? Once the observer is faced with a model which underwent a neglected experimental failure (that is, solved through authority-based emergency procedures), she will understand that "there are no actual, concrete systems in the world around us fitting the description it contains" [34, p. 283]. The experimental learning achieves the result of teaching scientific theories as something *necessarily* abstract and incoherent with everyday perceived reality: such a configuration of the experiment awkwardly clashes with Hacking's breakthrough intuition, according to which experiments (and the models they embed) *create* phenomena that might very well not give themselves in everyday reality [10]. The constructed/modeled nature of phenomena is a consequence of the experimental framework, and not something that the experiment must cope with as the byproduct of the clash between theory and actual reality.

4 Conclusion

The aim of this paper was to provide a sensible analysis of the veritable experimental framework in science. As noted in footnotes 5 and 17, this study is coherent with—and was partially inspired by—Magnani's conception of "epistemic warfare", which sees

> [...] scientific enterprise as a complicated struggle for rational knowledge in which it is crucial to distinguish epistemic (for example models) from non epistemic (for example fictions, falsities, propaganda, etc.) weapons. 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 [5] 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, for example scientific models, which 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 ancillary role in the epistemic warfare. So to say, these last aspects support scientific reasoning providing non-epistemic weapons able for example to persuade others 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 [19, p. 3].

Magnani's concept was devised arguing about the use and nature of models in science, but it can be applied fruitfully to the understanding of other aspects of scientific endeavor. Thinking of generative and demonstrative experiments, it can be said that the former reflect epistemic weaponry, while the latter partake of a non-epistemic nature. Yet both kinds of experiment are crucial and unremovable for a correct functioning of science: while generative experiment engage the natural framework, and are thus the first-line of scientific and technological progress, demonstrative experiments engage the human framework. Science is a human activity, therefore a fittingly shaped human framework (eager to invest funds, commitments, priorities etc.) is just as essential as the correct exercise of method and rationality.

The distinction I proposed should not be considered a dichotomy, but rather consists in the two poles of a continuum specter covering the experimental dimension. Even if it is possible to find some experiments (as in Newtonian mechanics) that are carried out only in patently demonstrative settings, there is not a fixed number of repetitions after which an experiment switches from being generative to demonstrative: Popper had already faced this problem, when dealing with the *diminishing returns* from repeated experiments [23].[18] On the other hand, the distinction between the two kinds of experiment is sometimes blurred in the actual scientific practice (not in the dissemination to a lay public): as shown by Ampère's

[18] See also footnote 14.

example in Sect. 3.1 (and other ones in [10]), what I called generative experiment has often had a scaffolding role, and once its outcome is assessed, the scaffolding is replaced by a more straightforward and *nicer* experiment informed by the already confirmed theory. Lastly, demonstrative experiments have a minor (if only nominally) role to play as *watchdogs* of the adequacy of well-assessed theories. That is to say—in Lakatosian terms?—they provide a further protective layer to the protective belt of a research programme: repeating *ad nauseam* experiments about basic chemical reaction, light properties, metal bars that expand when heated and so on, we keep assessing the adequacy of fundamental predictions.

It should be noted that even to consider the distinction as two poles of a continuum is slightly problematic because of some anomalies posed by contemporary sciences: in robotics, computer sciences or for instance genetics most experiments can be generative and demonstrative at the same time. A robot, for instance, is at once the product of the manipulative transformation generating new knowledge, and the mediator of dissemination of that same knowledge. This aspect is worth further studying, as is the relationship between my distinction and thought experiments: thought experiments can be seen at the same time as both generative and demonstrative experiments, depending on the conception of thought experiment rooted in one's background [2, 8]. If one considers thought experiments as reducible to arguments, then she might think of them as *demonstrative*; conversely if thought experiments are seen as rightful experiments, then—no matter how many times a thought experiment is repeated—it could remain *perennially generative*.

Acknowledgments I would like to thank the participants of the MBR'012_Italy conference for the valuable comments on my presentation: many of them contributed to the final form of this paper. In particular, I would like to thank Viola Schiaffonati for the insight about the anomaly of robotics with respect to my distinction.

References

1. Bertolotti, T.: From mindless modeling to scientific models: the case of emerging models. In: Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science: Western & Eastern Studies, pp. 75–104. Springer, Heidelberg (2012)
2. Bishop, M.A.: Why thought experiments are not arguments. Philos. Sci. **66**(4), 534–541 (1999)
3. Cetina, K.K.: Epistemic Cultures. How Sciences Make Knowledge. Harvard University Press, Cambridge (1999)
4. Dear, P.: Revolutionizing the Sciences: European Knowledge and its Ambitions, 1500–1700. Princeton University Press, Princeton (2001)
5. Feyerabend, P.: Against Method. Verso, London-New York (1975)
6. Gabbay, D.M., Woods, J.: The new logic. Log. J. IGPL **9**(2), 141–174 (2001)
7. Galison, P.: Philosophy in the laboratory. J. Philos. **85**(10), 525–527 (1988)
8. Gendler, T.S.: Galileo and the indispensability of scientific thought experiment. Br. J. Philos. Sci. **49**(9), 397–424 (1998)
9. Gooding, D.: Experiment and the Making of Meaning. Kluwer, Dordrecht (1990)

10. Hacking, I.: Representing and Intervening. Introductory Topics in the Philosophy of Natural Science. Cambridge University Press, Cambridge (1983)
11. Hacking, I.: On the stability of the laboratory sciences. J. Philos. **85**(10), 507–514 (1988)
12. Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago, 1962. Second expanded edition, 1970
13. Lakatos, I.: Proofs and Refutations. The Logic of Mathematical Discovery. Cambridge University Press, Cambridge (1976)
14. Latour, J.: Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge (1987)
15. Macdoland, S., Basu, P. (eds.): Exhibition Experiments. Blackwell, Malden (2007)
16. Magnani, L.: Thinking through doing, external representations in abductive reasoning. In: AISB 2002 Symposium on AI and Creativity in Arts and Science, London, Imperial College (2002)
17. Magnani, L.: Morality in a Technological World. Knowledge as Duty. Cambridge University Press, Cambridge (2007)
18. Magnani, L.: Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Springer, Berlin (2009)
19. Magnani, L.: Scientific models are not fictions: model-based science as epistemic warfare. In: Magnani, L., Li, P. (eds.) Philosophy and Cognitive Science: Western & Eastern Studies, pp. 1–38. Springer, Heidelberg (2012)
20. Mäki, U.: MISSing the world. Models as isolations and credible surrogate systems. Erkenntnis **70**, 29–43 (2009)
21. Mayo, D., Spanos, A.: Introduction and background. In: Mayo, D., Spanos, A. (eds.) Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, pp. 1–27. Cambridge University Press, Cambridge (2010)
22. Mayo, D.: Explanation and testing exchanges with clark glymour. In: Mayo, D., Spanos, A. (eds.) Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, pp. 351–363. Cambridge University Press, Cambridge (2010)
23. Musgrave, A.: Popper and 'diminiscing returns from repeated tests'. Australas. J. Philos. **53**(3), 248–253 (1975)
24. Musgrave, A.: Critical rationalism, explanation, and severe tests. In: Mayo, D., Spanos, A. (eds.) Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, pp. 88–112. Cambridge University Press, Cambridge (2010)
25. Nersessian, N.J., Patton, C.: Model-based reasoning in interdisciplinary engineering: cases from biomedical engineering research laboratories. In: Meijers, A.W.M. (ed.) The Handbook of the Philosophy of Technology and Engineering Sciences. Springer, Berlin/New York, Forthcoming (2009)
26. Nersessian, N.J.: How do scientists think? Capturing the dynamics of conceptual change in science. In: Giere, R.N. (ed.) Cognitive Models of Science, Minnesota Studies in the Philosophy of Science, pp. 3–44. University of Minnesota Press, Minneapolis (1992)
27. Nersessian, N.J.: Should physicists preach what they practice? Constructive modeling in doing and learning physics. Sci. Educ. **4**, 203–226 (1995)
28. Picha, M.: How to reconstruct a thought experiment. Organon F **18**(2), 154–188 (2011)
29. Popper, K.R.: The Logic of Scientific Discovery. Hutchinson, London, New York (1959)
30. Raftopoulos, A.: Is perception informationally encapsulated? The issue of theory-ladenness of perception. Cogn. Sci. **25**, 423–451 (2001)
31. Raichvarg, D.: Science on the fairgrounds: from black to white magic. doi: 10.1007/s11191-006-9011-4 (2006)
32. Schickore, J.: The significance of re-doing experiments: a contribution to historically informed methodology. Erkenntnis **75**, 325–347 (2011)

33. Steinle, F.: Entering new fields: exploratory uses of experimentation. Philos. Sci. **64**, S65–S74 (1996)

34. Thomson-Jones, M.: Missing systems and the face value practice. Synthese **172**, 283–299 (2010)

35. Woods, J.: Epistemic bubbles. In: Artemov, S., Barringer, H., Garcez, A., Lamb, L., Woods, J. (eds.) We Will Show Them: Essays in Honour of Dov Gabbay (Volume II), pp. 731–774. College Publications, London (2005)

Complex Systems of Mindful Entities: On Intention Recognition and Commitment

Luís Moniz Pereira, The Anh Han and Francisco C. Santos

Abstract The mechanisms of emergence and evolution of cooperation in populations of abstract individuals with diverse behavioural 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, 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 both Intention Recognition and to Commitment (separately and jointly), encompassing errors in decision-making and communication noise. As a result, both the emergence and stability of cooperation become reinforced comparatively to the absence of such cognitive abilities. This holds separately for Intention Recognition and for Commitment, and even more when they are engaged jointly. The present paper aims to sensitize the reader to these Evolutionary Game Theory based studies and issues, which are accruing in importance for the modelling of

L. M. Pereira (✉) · T. A. Han
Centro de Inteligência Artificial (CENTRIA), Departamento de Informática,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
2829-516 Caparica, Portugal
e-mail: lmp@fct.unl.pt

T. A. Han
e-mail: h.anh@ai.vub.ac.be

T. A. Han
AI-lab, Vrije Universiteit Brussel, Pleinlaan 2 1050 Brussels, Belgium

F. C. Santos
INESC-ID, Instituto Superior Técnico and ATP-group, Instituto para
a Investigação Interdisciplinar, Universidade Técnica de Lisboa, IST-Taguspark
2744-016 Porto Salvo, Portugal
e-mail: franciscocsantos@ist.utl.pt

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_28, © Springer-Verlag Berlin Heidelberg 2014

minds with machines, with impact on our understanding of the evolution of mutual tolerance and cooperation. In doing so, it also provides a coherent bird's-eye view of our own varied recent work, whose more technical details and results are spread throughout a number of well recognized publishing venues, and to which we refer the reader for a fuller support of our claims where felt necessary.

1 Introduction

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 behaviours 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 pre-defined but instead something like a shape, 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 intrinsic properties of elements [16].

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 [3, 37, 64, 94]. To understand the evolutionary mechanisms that promote and keep cooperative behaviour 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 macromolecules in a physical or biological context, and people, machines or organizations in a socioeconomic context.

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, the issue of self-consideration versus "the other"-consideration has fascinated thinkers, but the use of formal models and experimental games is relatively recent. Since Robert Trivers [104, 105] introduced the evolutionary approach to reciprocity, games have served as models to explore the issue. The modelling 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 repeatedly meet other individuals, each time doing so in randomized pairs, in a joint iterated game taking place within the scope of the whole population. 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) provides the means to understand the why and the how of what it takes for agents with individual interests to cooperate for a common weal [45, 64].

In its simplest form, a cooperative act is metaphorically described as the act of paying a cost to convey a benefit to someone else. If two players simultaneously decide to cooperate or not, the best possible response will be to try to receive the benefit without paying the cost. In an evolutionary setting, we may also wonder why would natural selection equip selfish individuals with altruistic tendencies while it incites competition between individuals and thus apparently rewards only selfish behaviour? Several mechanisms responsible for promoting cooperative behaviour have been recently identified [65, 94]. From kin and group ties [102, 111], to different forms of reciprocity [47, 66, 68, 72, 105] and networked populations [54, 87, 89, 90, 99], several aspects have been shown to play an important role in the emergence of cooperation.

Moreover, more complex 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 channelled to just 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 behaviours based on suggestive information about the intentions of their interaction partners—these in turn influenced by the behaviour of the individual himself—individuals are also influenced by their tolerance to error or noise in the communication. One hopes that, to start with, understanding these capabilities can be transformed into mechanisms for spontaneous organization and control of swarms of autonomous robotic agents [7], these being envisaged as large populations of agents where cooperation can emerge, but not necessarily to solve a priori given goals, as in distributed AI.

With these general objectives, we have specifically studied the way players' strategies adapt in populations involved in cooperation games. We used the techniques of EGT, considered games such as the Prisoner's Dilemma and Stag Hunt, and showed how the actors participating in repeated iterations in these games can benefit from having the ability to recognize the intentions of other actors, or to establish commitments, or both, thereby leading to an evolutionary stable increase in cooperation [27, 30–32], compared to extant best strategies.

Intention recognition (IR), or abducing intent, can be implemented using Bayesian Networks (BN) [30, 77, 78], taking into account the information of current signals of intent, as well as the mutual trust and tolerance accumulated from previous one-on-one play experience—including how my previous defections may influence another's intent—but without resorting to information gathered regarding players' overall reputation in the population. A player's present intent can be understood here as how he's going to play the next round with me, whether by cooperating or defecting. Intention recognition can also be learnt from a corpus of prior interactions among game strategies [31, 32], where each strategy can be envisaged and detected as players' (possibly changing) intent to behave in a certain way [28]. In both cases, we experimented with populations with different proportions of diverse strategies in order to calculate, in particular, what is the minimum fraction of individuals capable of intention recognition for cooperation to emerge, invade, prevail, and persist. It seems to us that even basic intention recognition, and its use in the scope of cooperation and tolerance, is a foundational cornerstone where we should and indeed began at, which was naturally followed by the capacity to establish and honour commitments [33, 34], as a tool towards the successive construction of collective intentions and social organization [92, 93].

We argue that the study of these issues in minds with machines has come of age and is ripe with research opportunities, and communicate below some of the published inroads we have achieved with respect to intention recognition, to commitment and to the emergence of cooperation, involving tolerance and intolerance, in the evolutionary game theory context.

2 Intention Recognition Promotes the Emergence of Cooperation

Most studies on the evolution of cooperation, grounded on evolutionary dynamics and game theory, have neglected the important role played by a basic form of intention recognition in behavioural evolution. In this section, we address explicitly this issue, characterizing the dynamics emerging from a population of intention recognizers. We derive a Bayesian Network model for intention recognition in the context of repeated social dilemmas and evolutionary game theory, by assessing the internal dynamics of mutual trust and tolerance, accumulated from previous one-on-one play experience, between intention recognizers and their opponents, as detailed below. Intention recognizers are then able to predict the next move of their opponents based on past direct interactions, which, in turn, enables them to prevail over the most famous strategies of repeated dilemmas of cooperation, even in the presence of noise. Overall, our framework offers new insights on the complexity and beauty of behavioural evolution driven by elementary forms of cognition.

2.1 Background

Intention recognition can be found abundantly in many kinds of interactions and communications, not only in human but also many other species [101]. The knowledge about intention of others in a situation could enable to plan in advance, either to secure a successful cooperation or to deal with potential hostile behaviours [27, 29, 83, 106]. Given the advantage of knowing the intentions of others and the abundance of intention recognition among different species, it is clear that intention recognition should be taken into account when studying or modeling collective behaviour. This issue becomes even more relevant when the achievement of a goal by an individual does not depend uniquely on its own actions, but also on the decisions and actions of others, namely when individuals cooperate or have to coordinate their actions to achieve a task, especially when the possibility of communication is limited [41, 52, 107]. For instance, in population-based artificial intelligence applications [1, 7, 26], such as collective robotics and others, the inherent problem of lack of intention recognition due to the simplicity of the agents is often solved by assuming homogeneous populations, in which each agent has a perfect image of the other as a copy of their own self. Yet, the problem remains in heterogeneous agent systems where it is likely that agents speak different languages, have different designs or different levels of intelligence; hence, intention recognition may be the only way agents understand each other to secure successful cooperation or coordination among heterogeneous agents. Moreover, in more realistic settings where deceiving may offer additional profits, individuals often attempt to hide their real intentions and make others believe in pretense ones [78, 82, 88, 98, 101].

Intention recognition is defined, in general terms, as the process of becoming aware of the intention of another agent and, more technically, as the problem of inferring an agent's intention through its actions and their effects on the environment [12, 41, 51]. For the recognition task, several issues can be raised grounded on the eventual distinction between the model an agent creates about himself and the one used to describe others, often addressed in the context of the "Theory of Mind" theory, which neurologically reposes in part on "mirror neurons", at several cortical levels, as supporting evidence [46, 61, 81].

The problem of intention recognition has been paid much attention in AI, Philosophy and Psychology for several decades [8, 9, 12, 21, 51]. Whereas intention recognition has been extensively studied in small scale interactive settings, there is an absolute lack of modelling research with respect to large scale social contexts; namely the evolutionary roles and aspects of intention recognition.

2.2 Modeling Behavioural Dynamics

Our study is carried out within the framework of Evolutionary Game Theory (EGT) [45, 58]. Here, individual success (or fitness) is expressed in terms of the outcome of a 2-person game, which, in turn, is used by individuals to copy others whenever these appear to be more successful. Comparative accumulated payoffs are used to update the population: more successful individuals produce more offspring, which inherit their strategy. Equivalently, the same process can be seen as if, instead of inheriting strategies, new individuals adapt by copying strategies from acquaintances that did better. Overall, this type of dynamics can be conveniently described as an ordinary differential equation—the replicator equation [45]—, which nicely describes any simple evolutionary process.

In our work we model intention recognition within the framework of repeated interactions. In the context of direct reciprocity [47, 48, 65, 105, 108] intention recognition is being performed using the information about past *direct* interactions. We study this issue using the well-known repeated Prisoner's Dilemma (PD) [95], i.e., so that intentions can be inferred from past individual experiences. Naturally, the same principles could be extended to cope with indirect information, as in indirect reciprocity [68, 69, 72]. This eventually introduces moral judgment and concern for individual reputation, which constitutes "per se" an important area where intention recognition may play a pivotal role [35, 72]. Here, however, we shall concentrate on the simpler case of intention recognition from past experiences.

Contrary to other approaches dealing with the integration of (direct or indirect) information about the past in individual decisions, e.g. in [57, 69, 109, 110], intention recognition is performed using a Bayesian Network (BN) model. BNs have proven to be one of the most successful approaches for intention recognition [12, 21, 29, 78, 100]. Their flexibility for representing probabilistic dependencies as well as causal relations, and the efficiency of inference methods have made them an extremely powerful tool for problem solving under uncertainty [73, 74], and appropriate to deal with several probabilistic as well as causal dependencies occurring in intention recognition. We derive a Bayesian Network model for intention recognition in the context of social dilemmas, taking into account mutual trusts between the intention recognizer and his opponent. Trusts are accumulated through past interactions, assuming that intention recognizers have a memory. Greater memory sizes enable to build longer-term mutual trusts, and therefore allow better tolerance to the errors of intended actions.

The repeated (or iterated) PD is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments [3, 4]. *TFT* starts by cooperating, and does whatever the opponent did in the previous round. It will cooperate if the opponent cooperated, and will defect if the opponent defected. But if there are erroneous moves because of noise (i.e. an intended move is wrongly performed with a given execution error, referred here as "noise"), the performance of *TFT* declines, in two ways: (i) it cannot correct errors and (ii) a population of *TFT* players is undermined by random drift when *AllC* (always cooperate) mutants appear (which allows exploiters to grow). Tit-for-tat is then advantageously replaced by generous tit-for-tat (GTFT), a strategy that cooperates if the opponent cooperated in the previous round, but sometimes cooperates even if the opponent defected (with a fixed probability $p > 0$). *GTFT* can correct mistakes, but remains suffering the random drift; in addition, it deals with pure defectors worse than *TFT*.

Subsequently, *TFT* and *GTFT* were replaced by win-stay-lose-shift (WSLS) as the winning strategy chosen by evolution [67]. *WSLS* repeats the previous move whenever it did well, but changes otherwise. *WSLS* corrects mistakes better than *GTFT* and does not suffer random drift. However, it is exploited seriously by pure defectors.

We consider a population of constant size N. At each evolution step, a random pair of players are chosen to play with each other. The population consists of pure cooperators, pure defectors plus either of *TFT*s or of *WSLS*s or of intention recognizers who, being capable of recognizing another's intention based on the past interactions, seek the cooperators to cooperate with and to defect toward detected defectors.

Interactions are modeled as symmetric two-player games defined by the payoff matrix, used by all players. In particular, each type of player chooses to play in the same way under the same circumstances.

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R,R & S,T \\ T,S & P,P \end{pmatrix} \end{array}$$

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [56, 87, 94]. Namely, in this work we are concerned with the Prisoner's Dilemma (PD), where $T > R > P > S$. In a single round, it is always best to defect, but cooperation may be rewarded if the game is repeated. In repeated PD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation.

Before providing our intention recognition model in the framework of social dilemmas, let us provide the definition of Bayesian Network. A Bayesian Network (BN) is a pair consisting of a directed acyclic graph (DAG) whose nodes represent variables and missing edges encode conditional independencies between the variables, and an associated probability distribution satisfying the Markov assumption of conditional independence, saying that variables are independent of non-descendants given their parents in the graph [73, 74].

In a BN, associated with each node of its DAG is a specification of the distribution of its variable, say A, conditioned on its parents in the graph (denoted by $pa(A)$)—i.e., $P(A|pa(A))$ is specified. If $p(A) = \emptyset$ (A is called root node), its unconditional probability distribution, $P(A)$, is specified. These distributions are called Conditional Probability Distribution (CPD) of the BN. The joint distribution of all node values can be determined as the product of conditional probabilities of the value of each node on its parents.

In [78], a general BN model for intention recognition is presented and justified based on Heinze's intentional model [41, 100]. Basically, the BN consists of three layers: cause/reason nodes in the first layer (called *pre-intentional*), connecting to intention nodes in the second one (called *intentional*), in turn connecting to action nodes in the third (called *activity*). Intuitively, the observed actions of an agent are causally affected by his/her intentions, which are in turn causally affected by the causes/reasons for which he committed to the intentions [8, 9]. The interested readers are referred to [41, 78, 100] for detailed discussions.

Based on this general model, we present an intention recognition model in the context of the social dilemmas, taking into account the past *direct* interactions (Fig. 1). The model is described from the view of an intention recognizer (denoted by $\mathcal{I}$) with respect to a co-player (denoted by $\mathcal{J}$), whose intention (C or D) is to be recognized. A player's intentions here can be understood as the characters or types of the player: how cooperative or defective he is in general when playing with me. Saying that the co-player has intention C (resp., D) means that, in general, he intends to cooperate with me (resp., exploit or defect towards me). Thus, if he has been cooperative in the past, it is likely he will continue to cooperate in the current interaction.

$\mathcal{J}$'s intention in a given interaction is causally affected by the trust he holds towards his opponent ($\mathcal{I}$), which is accumulated over their past (observed) interactions. $\mathcal{J}$'s intention in turn has given rise to his past actions. Let $M > 0$ be the memory size of intention recognizers, i.e. they can remember their moves and their opponents' moves in the last M rounds of interaction with any specific players.

For this Bayesian Network, we need to determine the prior probability of the node *oTrust*, i.e. $P(Tr)$; the CPD table of node *Intention*—specifying the conditional probability of $\mathcal{J}$ having an intention (C or D) given the trust he holds towards his opponent ($\mathcal{I}$), i.e. $P(I|Tr)$; and the CPD table of the node *pastObs*—specifying the conditional probability of the past observations given $\mathcal{J}$'s intention (C or D), i.e. $P(O|I)$.

The accumulated payoff from all interactions (not shown here) emulates the individual *fitness* or social *success* and the most successful individuals will tend to be imitated by others, implementing a simple form of social learning [24, 94, 103].

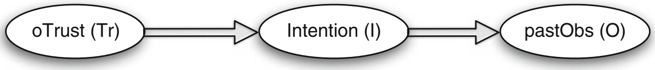


Fig. 1 Bayesian Network for Intention Recognition in Social Dilemmas. Pre-intentional level has one node, *oTrust (Tr)*, receives Boolean values, t (*true*) or f (*false*), representing the other's trust on us (the intention recognizers). Intentional level has one node, *Intention (I)*, receiving value C or D, corresponding to more cooperative and more defective, respectively, in the past. It is causally affected by *oTrust*. Activity level has one node, *pastObs (O)*, causally affected by *Intention* node. Its value is a pair (n_C, n_D) where n_C and n_D are the number of times the recognized player cooperated and defected, respectively, in the recent M (memory size) steps. *pastObs* is the only observed (evidence) node

Any player (including *IR*) can change its strategy by adopting another player's strategy with a probability defined by the Fermi distribution [64]. If a strategy has a higher (average) payoff or fitness than another, it tends to be imitated more by the other. The *IR* strategy in general has higher fitness than all others, thus it tends to by imitated by them, thereby dominating the population most of the time.

In the commonly used settings, including when interacting solely with the pure strategies (AllC and AllD) and when all considered strategies interacted with each other [47, 48, 65], IR always outperforms TFT and WSLS [30, 32]. The population spends more time in the homogeneous state of all being IRs, even in the presence of noise and of small random mutations. Furthermore, since a population of IRs is highly cooperative, it is clear that the introduction of intention recognition significantly increases the cooperation level of the population, leading to a greater social welfare.

2.3 Discussion

Using the tools of EGT, we have addressed the role played by intention recognition in the evolution of cooperation. In this work, we have shown, in a novel way, the role of intention recognition for the emergence of cooperation within the framework of the repeated Prisoner's Dilemma. Intention recognition is performed using a Bayesian Network model via computing mutual trusts between the intention recognizers and their opponents. Given the broad spectrum of problems which are addressed using this cooperative metaphor, our result indicates how intention recognition can be pivotal in social dynamics. We have shown that the intention recognition strategy prevails over the most successful existent strategies (TFT, WSLS) of the repeated PD, even when players have a very limited memory. IR deals with AllD better than TFT—the best known defector-dealer—, and corrects mistake better than WSLS—the best known mistake-corrector [65, 95]. As a result, a homogenous population of IRs has a higher level of cooperation than the ones of WSLSs and TFTs, resisting the invasion of other strategies.

In [47], it has been shown that in absence of noise, in a population of AllCs, AllDs and TFTs, the population spends most of the time in a homogeneous state of TFTs. However, as we have shown elsewhere, it is not the case if noise is present, especially under strong selection. In absence of noise, IR behaves the as well as TFT. Moreover, IRs are selected by evolution in the latter case where noise is present. We have shown that in a population of AllCs, AllDs and IRs, the population spends most of the time in homogeneous state of IRs in a broad range of scenarios and parameters, especially when the intensity of selection is strong. We have also exhibited experimentally that in a population where all the five strategies AllC, AllD, TFT, WSLS and IR are present, IRs still prevail most of the time. Therefore, together with the fact that IRs can correct mistakes better than WSLSs and TFTs, the presence of IRs would significantly increase the overall level of cooperation of the population.

Additionally, we have shown the role of a large memory size in recognizing/correcting errors, that is in recovering from ongoing persistent mutual defection that may result from a move announcement mistake, or from communication channel noise. Having a greater memory size allows to build longer-term mutual trusts/distrusts, and hence enables to better recognize erroneous moves. It then enables to better tolerate of a selfish act made by cooperative trustful individuals, and refuses to cooperate after an erroneous cooperation made by a defective untrustworthy ones. Indeed, intention recognition gives rise to an incipient mechanism of commitment formation, from which future behaviours may be assessed and trust bonds established. Overall, our work provides new insights on the complexity and beauty of behavioural evolution driven by basic, elementarily defined, forms of cognition.

3 Commitment Promotes the Emergence of Cooperation

Agents make commitments towards others, the promise to enact their play moves in a given manner, in order to influence others in a certain way, often by dismissing more profitable options. Most commitments depend on some incentive that is necessary to ensure that the action is in the agent's interest and thus, may be carried out to avoid eventual penalties [22]. The capacity for using commitment strategies effectively is so important that natural selection may have shaped specialized signaling capacities to make this possible [5, 15, 63, 82, 88, 98]. And it is believed to have an incidence on the emergence of morality [85]. Assuming cooperation to be, at best, just the result of individuals' purely competitive strategies can make it conceptually unstable [71], most especially in non-iterated or history-free interactions. And it seems possible that the spread of simplistic notions, rooted in science, about the evolutionary origins of social relationships could foster a trend to make these relationships more conflicted, and society more brutal. An antidote is an evolutionary approach to behaviour that incorporates a capacity for mutual commitment, shown advantageous for all concerned [63], even in non-iterated or memory-free settings.

Our goal is to examine, through EGT [45, 58, 94], how the most simple of commitment strategies work, and how they can give rise to the emergence of cooperation. We shall do so in the setting of the non-iterated Prisoner's Dilemma (PD).

In a nutshell, convincing others of one's credibility in a commitment proposal amounts to submit to options that change the incentives of the situation. These options, namely commitment cost and penalty for defaulting, can be expressed by the payoffs specified in a game. When opponent players observe meticulously such payoffs, and realize that compliance with a proposed commitment is in the proposing player's best interests, then, given any opponent player's open option to commit, these may change their expectations and behaviour accordingly, and adopt as a result a strategy which either accepts commitment proposals or ignores them. In general, there are four main reasons to believe a commitment will be fulfilled [63]: (i) a commitment can be self-reinforcing if it is secured by incentives intrinsic to the situation; (ii) a commitment can be secured by external incentives controlled by third parties; (iii) a commitment can be backed by a pledge of reputation; and (iv) a commitment can be reinforced by internal emotional motives. The first two types are secured in much the same way a loan is secured by a collateral. They objectively change the situation so that fulfillment becomes in the individual's best interests. The latter two types do not change the objective contingencies; they are subjective commitments in that they may involve a continued option of reneging, according to some or other stance extraneous to the game's given payoff matrix.

In this section, we provide a new EGT model showing that individuals tend to engage in commitments, which leads to the emergence of cooperation even without assuming repeated interactions. The model is characterized by two key parameters: the punishment cost of failing commitment imposed on either side of a commitment, and the cost of managing the commitment deal. Our analytical results and extensive computer simulations show that cooperation can emerge if the punishment cost is large enough compared to the management cost.

3.1 Model

In our EGT setting, the game's payoff matrix summarily ingrains and expresses in its structure the impingement of all such contingencies [33, 34]. For instance, often a capacity for commitment allows individuals to act in ways that reap the benefits of image scoring through maintaining a reputation, or the access of others to a social history of prior interactions. In this study, for simplicity but also for exhibiting the purity and power of the commitment mechanism, we ignore the effect of repeated interactions [104, 105], and of any reputation [68, 72] associated with particular individuals. We have shown [33, 34] that the simplest of core commitment mechanisms can improve cooperation, and leave any other complications for the future, most promisingly how commitment can be combined with and reinforce other known mechanisms of cooperation, for instance, intention recognition [30–32]. And perhaps surprisingly we can do so. Thus, no credibility of commitment is taken into account [11] beyond that which is expressed in a game's payoff matrix. No reputation appraisal of the commitment proposer is made by its co-player, and no historical or social data is even available to do so. Each pairwise interaction is purely based on fixed individual strategies that might involve commitment or the lack thereof. Also, no cheater or deceit detection or intention recognition is in place [30, 31, 49]. Nevertheless, systematic unconditional bluffing on the part of a player is a possible fixed feature of its strategy, in the sense that, from the start, the player does not intend to fulfill commitments. In our commitment model players defaulting on their commitments, be they the proposing or the accepting party, are subject to evolutionary disadvantage for a wide range of parameters. Commitments come at a price: players must pay to propose commitment, but commitment acceptors that default are penalized a compensation value in favor of the proposer. We have shown, with the model below, that more elaborate commitment strategies are not strictly necessary for commitment to become evolutionarily advantageous. Neither an aptitude for higher cognition, nor for empathy, nor for mind reading are needed. These aptitudes would only be required for more sophisticated forms of commitment, scaffolded atop the core one. We have explained the evolution, in a population, of the capacity for a simple form of commitment as the result of otherwise being excluded from a group of committed promise abiding cooperators, in the sense that this strategy tends to invade the game playing population under rather general conditions.

Let us consider a commitment variant of the Prisoner's Dilemma game in which a new type of cooperator (denoted by COM_C) that, before each interaction, asks the co-player whether it commits to cooperate. If the co-player does not so commit, there is no interaction. Both players get 0. Otherwise, if the co-player commits, they then go on to play with each other in the present interaction. If the co-player keeps to its commitment, both players obtain the reward payoff, R. Otherwise (if the co-player fails its commitment), the proposing or focal player obtains the sucker payoff, S, and its co-player obtains the temptation payoff, T. However, the one that fails the commitment will suffer a penalty cost, and its non-defaulting co-player gains a compensation for the potential loss due to its default of fulfilling the commitment. For simplicity, we assume that these two amounts (penalty and compensation) are equal, being denoted by δ. The penalty cost can be a real monetary one, e.g., in the form of prior debit (e.g., in the case of accommodation rental) or of a subsequent punishment cost (e.g., commitment was performed in terms of a legal contract, and one who fails commitment must pay a cost to compensate for the other), or an imaginary abstract value, e.g., public spread of good/bad reputation (bad reputation for the one that fails, and sympathy for the other), or even an emotional suffering [22, 43, 63, 85]. How this cost is set up depends on the types of commitment at work, or the reason for which the commitment is believed to be fulfilled (see beginning of Sect. 3), which topic is beyond the scope of this paper. However, various techniques can be seen in [43, 91].

Two players that defect in an interaction obtain the punishment payoff, P. For setting up a commitment, the proposer must pay a small management cost, ε. The cost of proposing and setting up the commitment might be high, but it is reasonable to assume that this cost is quite small compared to the mutual benefit of a cooperation strategy guaranteeing commitment, $\varepsilon < < R$.

We consider a finite population of a constant size, consisting of four strategies: COM_C (as described above), C (always cooperates, without proposing to commit), D (always defects, and does not commit when being asked to), and D_COM (always defects, though commits when being asked to). Here, for the sake of exposition, we assume that cooperators, including COM_C and C players, always commit whenever being asked to since they are better off to do so, as cooperation is their default choice, and reasonable commitment deals only are proposed.

In each round, two random players are chosen from the population for an interaction. For the row player, the (average) payoff matrix is consequently rendered as:

$$\begin{array}{c c} & \begin{array}{cccc} COMC & C & D & DCOM \end{array} \\ \begin{array}{c} COMC \\ C \\ D \\ DCOM \end{array} & \begin{pmatrix} R-\varepsilon/2 & R-\varepsilon & -\varepsilon & S+\delta-\varepsilon \\ R & R & S & S \\ 0 & T & P & P \\ T-\delta & T & P & P \end{pmatrix} \end{array} \tag{1}$$

Note that when a COM_C interacts with another COM_C, only one of them pays the cost of having proposed commitment, ε (e.g., the arbitrary one that proposes). Therefore, the average payoff of a COM_C in playing with another COM_C is, $R - \varepsilon/2$.

All in all, our study exhibits that, in spite of the absence of repeated interactions, reputation effect, network reciprocity, as well as group and kin selection, the strategy of commitment proposal may enable the emergence of cooperation, even under the presence of noise. By imposing a high cost for failing a commitment, when compared to the cost of setting up or managing the commitment deal, the commitment cooperative agents COM_C can get rid of the fake committers (D_COM) as well as avoid being exploited by the pure defectors (D), while playing approximately equally well against the pure cooperators (C). The results of this study suggest that our specialized capacity for commitment, which might have been shaped by natural selection [63], consists in a capacity for managing to impose a high cost of punishment, whether it is monetary or of abstract emotional or reputation value, with a relatively small cost.

Furthermore, the analytical results, supported by extensive computer simulations, showing the explicit relationships between the factors involved in the commitment mechanism would clearly provide important insight into the design of multi-agent systems resorting to commitments to facilitate cooperation [11, 13, 38, 43, 52, 91, 112, 113].

3.2 Related Work

Evolutionary explanations of commitment, particularly its role in the evolution of cooperation, have been actively sought for and discussed in several fields, including Psychology and Philosophy [5, 11, 15, 22, 43, 63, 85]. But there are only a few computational models that show the evolutionary advantages of commitment in problems where cooperative acts are beneficial [82, 88, 98]. In addition, often models rely on repeated interactions or long-term relationships [5, 15], alike the conditions where Triver's direct reciprocity [105] may play a role. Here we provide an analytic model in the framework of evolutionary game theory showing that, with the availability of the mechanism of commitment, cooperation can emerge even without assuming repeated interactions, or the availability of player reputation.

We note that there is a significant difference between our commitment model and works by others on costly punishment [17, 18, 40, 70, 80]. A commitment deal must be agreed by both sides of it in advance, thereby giving credibility and justification to punish any defaulting player. In addition, the prior agreement gives rise to compensation—the amount of which, in some cases, is agreed explicitly in advance—to the non-defaulting player. This compensation for the non-defaulting player is the significant difference that makes successful those players using the commitment strategy, while those using the costly punishment strategy have only a narrow margin of efficiency [70]; does not stand out as a winning strategy [17]; nor does it promote cooperation at all when taking into account antisocial punishment [42, 80]. The compensation might bring benefit to the commitment strategists once an appropriate deal would be arranged. This suggests that although costly punishment, whether it is social or antisocial, might not promote the evolution of cooperation, what we call ‘justified’ punishment, which is warranted by an appropriate commitment deal, does, so that bluffing committers are in the limit scourged. This kind of punishment might not be costly at all, and can even bring net benefit to its upholder, hence leading to the emergence of cooperation.

Last but not least, it is undoubtedly important to mention the extensive literature of AI and multi-agent systems research on commitment, e.g., [11, 13, 38, 43, 52, 91, 112, 113]. The main concern therein is how to formalize different aspects of commitment and how a commitment mechanism can be implemented in multi-agent interactions to enhance them (e.g. for improved collaborative problem solving [113]), especially in the context of game theory. In contradistinction, our concern is in the nature of an evolutionary explanation of commitment, particularly how it can promote the emergence of cooperation. More importantly, our evolutionary study of the commitment mechanism leads to insights about the global influence of the mechanism within a (large) population of agents, thereby enabling improvement for the design of multi-agent systems operating upon commitments [13, 112, 113].

3.3 Discussion

Within the general game theory concept of commitment, or intention manifestation, several distinctions can help separate different subtypes. In particular, some commitments are upfront promises of a next move that can help, while others are upfront threats of a subsequent move that can harm. Commitments can be conditional or unconditional. Threats are usually attempts to influence another person’s next move by stating a conditional subsequent move, and that’s how we may envisage them. Promises are more likely to be unconditional, and that’s how we may conceive of them, though more generally they can be conditional on the other fulfilling a matching promise.

Commitments can also be just towards oneself, taking into account the evolution of possible futures afforded by actions and events, and the individual's prior and post preferences, in what might be classically seen as a game against nature.

In [75, 76], three different types of individual commitment—hard, revocable, and momentary—are studied in such an evolution context. Let us recall that commitment, in the context of game theory, is a device or mechanism to decide the outcome with the other party [91]. Schelling distinguishes between commitment pure and simple and commitment that takes the form of a threat. What he calls "ordinary" commitment corresponds, in game theory, to the making of an opening announcement in a sequential play, which we dub preemptive, just before both players make their actual move. To constitute a preemption, a player's announcement action must be irrevocable, that is a promise that is assuredly kept. Preemptive commitment is not necessarily profitable, because it hinges on the opponent's actual move. Schelling however does not assume the other type of commitment as a "threat", which pertains to a player's move in reaction to the opponent's move. Threats, being conditional, may be of the "if-then-else" form, and can thus combine a threat and a promise, the latter albeit implicit whenever there are just two possible moves. We prefer instead to label "reactive" such so-called threat commitments. In the game context, these occur when the player with the last move irrevocably pledges to respond, in a specified but contingent way, to the opponent's prior choice [44].

In a nutshell, some players can be "preemptive" committers—those that always propose and always accept proposed commitments—, others may be "reactive" committers—those that always make a "reactive" statement and comply with the implicit requests in such statements—, while other players, though accepting to commit nevertheless default on their commitment, and even others simply omit and ignore preemptive or reactive commitments in their strategies—they might for instance be persistent defectors or persistent cooperators as we have seen, or, for that matter, follow any other strategy ignorant of commitment. Moreover, in iterated games, commitments can concern future rounds and not just the present one.

We purport to have shown that a simple commitment abiding cooperative strategy can be evolutionarily advantageous even in a non-iterated game setting. But much remains to be explored. In the more general setting and to avoid confusion, it can be helpful to distinguish, even if only conceptually, between "execution moves" and "pre-play moves" [44]. The terms first move and last move then always refer exclusively to execution moves—the choices that actually generate the payoffs. In contrast, commitments come earlier with respect to execution moves: they are pre-play moves. A preemptive commitment is a pre-play move that allows the player making it to take the first execution move. A reactive commitment, although also a pre-play move, can be made only by the player who has the last execution move. In either case, by giving up on his or her choice through committing, the commitment player leaves the opponent with "the last clear chance to decide the outcome" [91].

In our present game setting, however, there was no need to make the distinction between the first and the second to play, because each possible player strategy move is exhibited and fixed from the start, as expressed and ingrained in the payoff matrix. By so introducing the several committed unconditional move strategies—though the payoff is of course conditional on the opponent's move—, we can emulate what would happen in a round if a move sequence actually existed. Put briefly, our commitment model is of the simplest kind and, moreover, it is brought to bear solely on the very next move fold of a pair of players, with no history available on prior commitments. Nevertheless, it captures core features of commitment, namely the high cost of defaulting to discourage false commitment, and thus make it plausible, and a comparatively small but non-zero cost of commitment proposal to lend it initial credibility. On top of this core model more elaborate models affording commitment can subsequently be rooted, including those involving delayed deceit.

What's more, commitment (or intention manifestation) and intention recognition, are but two sides of a coin really, and their future joint study in the EGT setting is all but unavoidable [27]. It has become increasingly obvious that maximizing reproductive success often requires keeping promises and fulfilling threats, even when that requires in turn sacrifices regarding individual short-term interests. That natural selection has shaped special mental capacities to make this possible seems likely, including a capacity for commitment [63] and for intention recognition [30, 31]. The commitment stance goes yet further, and many aspects of human groups seem shaped by effects of commitments and intention recognition, namely group boundaries, initiation rituals, ideologies, and signals of loyalty to the group [96–98]. Conversely, many aspects of groups seem to exist largely to facilitate commitment to cooperate and to limit the utility of coercive threats.

The generalized ability for commitment to support cooperative interaction is an important aspect of plasticity in human behaviour, and humans support their deal-making in lots of ways. The law is full of instances of people using techniques of commitment to establish the honesty of their intentions, namely through a variety of contracts [23]. Institutions themselves are supported on committal contracts, and the law of the land proffers methods for constituting and of accountability of social institutions [93].

Given our rigorous approach's inroad results, we believe they lend promise to that further studies of commitment will benefit greatly from rigorous models that allow for their analytical study and computer simulation, and in particular within the fold of EGT for the better to examine the emergence of complex social behaviour.

4 Intention Recognition, Commitment, and Evolution of Cooperation

Individuals make commitments towards others in order to influence others to behave in certain ways. Most commitments may depend on some incentive that is required to ensure that the action is in the agent's best interest and thus, should be carried out to avoid eventual penalties. Similarly, individuals may ground their decision on an accurate assessment of the intentions of others. Hence, both commitments and intention recognition go side by side in behavioural evolution. Here, we analyze the role played by the co-evolution of intention recognition plus the emergence of commitments, in the framework of the evolution of cooperative behaviour. We resort to tools of evolutionary game theory in finite populations, showing how the combination of these two aspects of human behaviour can enhance the emergent fraction of cooperative acts under a broad spectrum of configurations.

There are cases where it is difficult, if not impossible, to recognize the intentions of another agent. It might be your first interaction with someone in your life, and you have no information about him/her which can be used for intention recognition. You also might know someone well, but you still might have very little relevant information in a given situation to predict the intentions with high enough confidence. Furthermore, you might also have abundance of relevant observations about him/her, but he/she is so unpredictable that you have rarely managed to predict his/her true intention in the past. In all such cases, the strategy of proposing commitment, or intention manifestation, can help to impose or clarify the intentions of others. Note that *intention is choice with commitment* [8, 14, 84]. Once an agent intends to do something, it must settle on some state of affairs for which to aim, because of its resource limitation and in order to coordinate its future actions. Deciding what to do established a form of commitment [14, 84]. Proposing a commitment deal to another agent consists in asking it to express or clarify its intentions.

One of the commitments we all know is marriage. By giving up the option to leave someone else, spouses gain security and an opportunity for a much deeper relationship that would be impossible otherwise [20, 63], as it might be risky to assume a partner's intention of staying faithful without the commitment of marriage. Though suggestive, this simplistic view of marriage also reveals some of the simplifications of the model. A marriage is indeed a commitment between partners. However, it is also a signal to the social group of the partners? cohesion with the group, and a signal that each partner sends to himself or herself, validating the choice of staying in the relationship.

A contract is another popular kind of commitment, e.g. for an apartment lease [20]. When it is risky to assume another agent's intention of being cooperative, arranging an appropriate contract provides incentives for cooperation. However, for example in accommodation rental, a contract is not necessary when the cooperative intention is of high certainty, e.g. when the business affair is between close friends or relatives.

Having said this, arranging a commitment deal can be useful to encourage cooperation whenever intention recognition is difficult, or cannot be performed with sufficiently high certainty. On the other hand, arranging commitments is not free, and requires a specific capacity to set it up within a reasonable cost (for the agent to actually benefit from it) [62, 63]—therefore it should be avoided when opportune. In the case of marriage, partners sometimes choose to stay together without an official commitment when it might be too costly (e.g., it could be against parents' or families' wish, or it may need to be in secret because of their jobs) and/or they strongly trust each other's faithfulness (e.g., because of emotional attachment [19, 20]). In short, a combination of the two strategies, those of commitment and of intention recognition, seems unavoidable. Nevertheless, intention recognition without actual commitment can be enhanced by costly engagement gifts, in support of sexual selection and attachment [39, 60]. Furthermore, social emotions can act as ersatz commitment [19].

Here, we start from the model [33] of commitment formation (described in the previous section), characterized by two key parameters: a punishment cost of failing commitment imposed on either side of a commitment deal, and the cost of managing it. On top of that model, again using EGT, we show that combining intention recognition and commitment strategies in a reasonable way can lead to the emergence of improved cooperation, not able to be achieved solely by either strategy. Our study seeks what is a reasonable combination of commitment and intention recognition.

We shall do so in the setting of the Prisoner's Dilemma (PD). It will be seen from our model that, in most of the cases, there is a wide range of combination of the intention recognition and commitment strategies, which leads to a strategy that performs better than either strategy solely—in the sense that the population spends more time in the homogeneous state of agents using that strategy [40, 47]. Our results suggest that, if one can recognize intentions of others with high enough confidence or certainty, one should rely more on it, especially when it is difficult to reach to a conceivably strong commitment deal. It helps to avoid the unnecessary cost of arranging and managing the deal. That is, in a transparent world where people have nothing to hide from each other, contracts are unnecessary.

On the other hand, when intention recognition with high precision is difficult (due to, e.g. environment noise, agents have great incentives to hide intentions, or there is not enough observed actions), one should rely more on the commitment strategy, particularly if a reasonable deal can be envisaged.

4.1 A Minimal Model Combining Intention Recognition and Commitment

We provide a new strategy, IRCOM, which combines the two strategies, those of intention recognition and commitment. In an interaction, IRCOM recognizes the intention (cooperates or defects) of its co-player [30]. A confidence level, *cl*, is assigned to the recognition result. It defines the degree of confidence (here in terms of probability) that IRCOM predicts the co-player's intention correctly.

Note that in AI the problem of intention recognition has been paid much attention for several decades, and the main stream is that of probabilistic approaches [2, 6, 10, 12, 41]. They tackle the problem by assigning probabilities to conceivable intentions (conditional on the current observations), based on which the intentions are ranked. Similarly to [2, 6, 28], in our model, a degree of confidence, cl, in terms of a probability measure, is assigned to intentions.

In general, cl follows some probability distribution. As in a real intention recognition problem, the distribution should depend on the intention recognition method at work (how efficient it is), the environment IRCOM lives in (is it supportive for gathering relevant information for the recognition process, e.g. observability of co-players' direct and indirect interactions, perception noise, population structure), etc. For example, we can consider different distributions satisfying that the longer IRCOM survives, the more precisely or confidently it performs intention recognition; or, considering the repeated interaction setting in the framework of the iterated PD, the more IRCOM interacts with its co-player, the better it can recognize the co-player's intention (see intention recognition models for the iterated PD in [30–32]).

We model cl by a continuous random variable X with probability density function $f(x, U)$, where U is a vector characterizing the factors that might influence cl, including the efficiency of the intention recognition model at work, the environmental factors (e.g., noise, population structure), and the interaction setting (repeated, one-shot, etc.).

If IRCOM is confident enough about the intention recognition process and result, that is cl is greater than a given, so-called, *confidence threshold* $\theta \in [0, 1]$, then in the current interaction IRCOM cooperates if the recognized intention of the co-player is to cooperate, and defects otherwise. The prediction is wrong with probability $(1 - cl)$. For simplicity, we assume that the prediction is a (continuous) random variable, Y, uniformly distributed in $[0, 1]$. Hence, the probability that IRCOM utilizes intention recognition, but with an incorrect and correct prediction, respectively, can be written as joint probability distributions [25, 36].

If $cl \leq \theta$, i.e. IRCOM is not confident enough about its intention prediction, it behaves the same as COM_C (see above). The greater θ is, the more cautious IRCOM is about its intention recognition result. Obviously, if $\theta = 1$, IRCOM behaves identically to COM_C ; and if $\theta = 0$, IRCOM behaves identically to a (pure) intention recognizer [30, 31].

We now replace COM_C with IRCOM, considering a population of four strategies, IRCOM, C, D, and D_COM. For the row player, the (average) payoff matrix reads $M = \theta M_1 + M_2$, where M_2 is the payoff matrix when IRCOM utilizes the intention recognition strategy, i.e. in the case $cl > \theta$. To derive M_2, we consider the case that cl has a uniform distribution in the interval $[0, 1]$, i.e. $f(x, U) = 1$ for $x \in [0, 1]$ and 0 otherwise.

The main subject of our published analysis is to address, given the payoff entries of the PD, and the parameters of the commitment deal IRCOM can manage, how confident about the intention recognition result IRCOM should be in order to make a decision, without relying on the commitment proposing strategy. That is, if there is an optimal value of θ for an IRCOM to gain greatest net benefit.

The results show that, whenever the intention recognition model is efficient enough, the intention recognition strategy solely (i.e. IRCOM with $\theta = 0$)

performs quite well, complying with the results obtained in [30–32], where concrete intention recognition models are deployed.

However, when a quite strong commitment deal can be envisaged, arranging it can still glean some evolutionary advantage. But in case only weak commitment deals can be arranged, it is then more beneficial to rely, even exclusively, on the intention recognition strategy, should it be efficient enough.

4.2 Discussion

A general implication of our analysis is that an appropriate combination of the two strategies of commitment and intention recognition often leads to a strategy that performs better than either one solely. It is advantageous to rely on the intention recognition strategy (when reaching sufficiently high confidence about its result) because it helps to avoid the cost of arranging and managing commitment deals, especially when no strong deals can be arranged or envisaged.

This result has a similar implication to that obtained in [50], where the authors show that overconfidence might give evolutionary advantage to its holders. In our model, an IRCOM can gain extra net benefit if it is a little overconfident (that is, when using sufficiently small θ), taking risk to rely on intention recognition result instead of arranging some commitment deal. Differently, because in our model IRCOM is further guaranteed by an efficient strategy of commitment, being over-overconfident (that is, using too small θ) and relying exclusively on intention recognition might prevent it from opportunely gaining benefit from the commitment strategy—especially in case the intention recognition model at work is not efficient. It said, the performance of overconfident individuals [50] can be enhanced by relying on the commitment strategy when they need to muster overly high courage (say, in order to decide to claim some resource).

In the framework where intention recognition is difficult and of high risk, for example, climate change negotiation [59, 79, 86], military setting—comprising a lot of bluffing [53, 91]—and international relationships [55], our model suggests arranging a strong commitment deal.

4.3 Conclusions

Assume simply that we are given an intention recognition method, that affords us a degree of confidence distribution *cl* about its predictions, with regard to the intentions of others, and hence their future actions, typically on the basis of their seen actions and surrounding historical and present circumstances. Assume too some commitment model is given us about providing mutual assurances, and involving an initial cost and a penalty for defaulting.

We have shown how to combine together one such general intention recognition method, with a specific commitment model defined for playing the Prisoner's Dilemma (PD), in the setting of Evolutionary Game Theory (EGT), by means of a single payoff matrix extended with a new kind of player, IRCOM, which chooses whether to go by the result of its intention recognition method about a co-player's next move, or to play by the commitment strategy, depending on whether its level of confidence on the intention prediction cl exceeds or not some a given confidence threshold θ. Our results indicate that IRCOM is selected by evolution for a broad range of parameters and confidence thresholds.

Then we have studied, for a variety of cl and θ, in the context of PD in EGT, how IRCOM performs in the presence of other well-known non-committing strategies (always cooperate, C, and always defect, D) – plus the strategy that commits when being asked to, but always defects, D_COM. Analytical and simulation results show under which circumstances, for different cl and θ, and distinct management and punishment costs, ε and δ, does the new combined strategy IRCOM prove advantageous and to what degree. And does indeed, IRCOM proves to be adaptably advantageous over those other just mentioned strategies and in all circumstances from a quite small confidence level onwards.

Much remains to be done with respect to further consideration of combining the two strategies of intention recognition and commitment. The two go often together, and not just in the basic way we have examined. Actually they are two sides of one same coin, one side being an attempt to identify an intention, the other being the manifestation of an intention. For one, we only considered the case where intention recognition comes first in order to decide on a commitment proposal. But, in general, once a commitment is made, intention recognition is a paramount method to follow up on whether the commitment will be honoured, on the basis of detecting or otherwise not the intermediate actions leading up to commitment fulfillment. Furthermore, the information about commitments can be used to enhance intention recognition itself.

It seems to us that intention recognition, and its use in the scope of commitment, is a foundational cornerstone where we should begin at, naturally followed by the capacity to establish and honour commitments, as a tool towards the successive construction of collective intentions and social organization [92, 93]. Finally, one hopes that understanding these capabilities can be useful in the design of efficient self-organized and distributed engineering applications [7], from bio- and socio-inspired computational algorithms, to swarms of autonomous robotic agents.

5 Coda

Evolutionary Psychology and Evolutionary Game Theory provide a theoretical and experimental framework for the study of social exchanges.

Recognition of someone's intentions, which may include imagining the recognition others have of our own intentions, and may comprise not just some error tolerance, but also a penalty for unfulfilled commitment, can lead to evolutionary stable win/win equilibriums within groups of individuals, and perhaps amongst groups. The recognition and the manifestation of intentions, plus the assumption of commitment—even whilst paying a cost for putting it in place—, are all facilitators in that respect, each of them singly and, above all, in collusion.

What is more, by means of joint objectives under commitment, one might promote the inclusion of heretofore separate groups into more global ones. The overcoming of intolerance shall benefit from both levels of manifest interaction—individual and group-wise.

We have argued that the study of these issues, of minds as evolving machines, has come of age and is ripe with research opportunities—including epistemological—and have communicated in some detail here some of the inroads we have explored, and have pointed to the much more detailed published results of what we have achieved, with respect to intention recognition, commitment, and mutual tolerance, within the overarching evolutionary game theory context.

The work of many other authors has also been emphasized and been given references, so the interested reader may easily begin to delve into this fascinating area, and follow up on its very active ongoing exploration and applications potential.

Acknowledgments This article is in great part a summary and cursory overview of the authors' detailed joint work mentioned in the publications. TAH acknowledges support from FCT-Portugal (grant SFRH/BD/62373/2009) and the FWO Vlaanderen (Postdoctoral fellowship). FCS acknowledges support from FCT-Portugal (PTDC/FIS/101248/2008, PTDC/MAT/122897/2010 and INESC-ID's PEst-OE/EEI/LA0021/2011). We thank all three reviewers for their very helpful indications.

References

1. Ampatzis, C., Tuci, E., Trianni, V., Dorigo, M.: Evolution of signaling in a multi-robot system: categorization and communication. Adapt. Behav. **16**(1), 5–26 (2008)
2. Armentano, M.G., Amandi, A.: Goal recognition with variable-order markov models. In: Proceedings of the 21st International Joint Conference on, Artificial Intelligence, pp. 1635–1640 (2009)
3. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984). ISBN 0-465-02122-2
4. Axelrod, R.: The evolution of cooperation. Science **211**, 1390–1396 (1981)
5. Back, Istvan, Flache, Andreas: The adaptive rationality of interpersonal commitment. Ration. Soc. **20**(1), 65–83 (2008)
6. Blaylock, N., Allen, J.: Statistical goal parameter recognition. In: Zilberstein S., Koehler J., Koenig S. (eds.) Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS'04), pp. 297–304. AAAI (2004)
7. Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, USA (1999)
8. Bratman, M.E.: Intention, Plans, and Practical Reason. The David Hume Series, CSLI (1987)
9. Bratman, M.E.: Faces of Intention: Selected Essays on Intention and Agency. Cambridge University Press (1999)

10. Bui, H., Venkatesh, S., West, G.: Policy recognition in the abstract hidden markov model. J. Artif. Intell. Res. **17**, 451–499 (2002)
11. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model (Wiley Series in Agent Technology). Wiley (2010)
12. Charniak, E., Goldman, R.P.: A Bayesian model of plan recognition. Artif. Intell. **64**(1), 53–79 (1993)
13. Chopra, A.K., Singh, M.P.: Multiagent commitment alignment. In: Proceedings of the 8th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS '09), pp. 937–944 (2009)
14. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artif. Intell. **42**(2–3), 213–261 (1990)
15. de Vos, H., Smaniotto, R.: Reciprocal altruism under conditions of partner selection. Ration. Soc. **13**(2), 139–183 (2001)
16. 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, MA (2003)
17. Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A.: Winners don't punish. Nature **452**(7185), 348–351 (2008)
18. Fehr, E., Gachter, S.: Altruistic punishment in humans. Nature **415**, 137–140 (2002)
19. Frank, R.H.: Passions Within Reason: The Strategic Role of the Emotions. W. W. Norton and Company, New York (1988)
20. Frank, Robert H.: Cooperation through emotional commitment. In: Nesse, R.M. (ed.) Evolution and the Capacity for Commitment, pp. 55–76. Russell Sage, New York (2001)
21. Geib, C.W., Goldman, R.P.: A probabilistic plan recognition algorithm based on plan tree grammars. Artif. Intell. **173**(2009), 1101–1132 (2009)
22. Gintis, H.: Beyond selfishness in modeling human behavior. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment. Russell Sage, New York (2001)
23. Goodenough, O.R.: Law and the biology of commitment. In: Nesse, R.M. (ed.) Evolution and the Capacity for Commitment, pp. 262–291. Russell Sage, New York (2001)
24. Szabó, G., Tőke, C.: Evolutionary prisoner's dilemma game on a square lattice. Phys. Rev. E **58**, 69–73 (1998)
25. Gut, A.: An Intermediate Course in Probability, 2nd edn. Springer Publishing Company, Incorporated, New York (2009)
26. Gutierrez, A., Campo, A., Santos, F.C., Monasterio-Huelin, F., Dorigo, M.: Social odometry: imitation based odometry in collective robotics. I. J. Adv. Robot. Syst. **2**(6), 129–136 (2009)
27. Han, T.A.: Intention recognition, commitments and their roles in the evolution of cooperation. Ph.D. thesis, Department of Informatics, Faculty of Sciences and Technology, Universidade Nova de Lisboa (May 2012)
28. Han, T.A., Pereira, L.M.: Context-dependent incremental intention recognition through Bayesian network model construction. In: Nicholson, A. (ed.) Proceedings of the Eighth UAI Bayesian Modeling Applications Workshop (UAI-AW 2011), vol. 818, pp. 50–58. CEUR Workshop Proceedings (2011)
29. Han, T.A., Pereira, L.M.: Intention-based decision making via intention recognition and its applications. In: Guesgen, H., Marsland, S. (eds.) Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security. IGI Global, (forthcoming) (2013)
30. Han, T.A., Pereira, L.M., Santos, F.C.: Intention recognition promotes the emergence of cooperation. Adapt. Behav. **19**(3), 264–279 (2011)
31. Han, T.A., Pereira, L.M., Santos, F.C.: The role of intention recognition in the evolution of cooperative behavior. In: Walsh, T. (ed.) Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'2011), pp. 1684–1689. AAAI (2011)
32. Han, T.A., Pereira, L.M., Santos, F.C.: Corpus-based intention recognition in cooperation dilemmas. Artif. Life j. **18**(4), 365–383 (2012)

33. Han, T.A., Pereira, L.M., Santos, F.C.: The emergence of commitments and cooperation. In: Proceedings of the 11th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'2012), pp. 559–566. ACM (2012)
34. Han, T.A., Pereira, L.M., Santos, F.C.: Intention recognition, commitment, and the evolution of cooperation. In: Proceedings of IEEE Congress on Evolutionary Computation, pp. 1–8. IEEE Press, June 2012
35. Han, T.A., Saptawijaya, A., Pereira, L.M.: Moral reasoning under uncertainty. In: Proceedings of the 18th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-18), pp. 212–227. Springer LNAI 7180 (2012)
36. Han, T.A., Traulsen, A., Gokhale, C.S.: On equilibrium properties of evolutionary multiplayer games with random payoff matrices. Theor. Popul. Biol. **81**(4), 264–272 (June 2012)
37. Hardin, G.: The tragedy of the commons. Science **162**, 1243–1248 (1968)
38. Harrenstein, P., Brandt, F., Fischer, F.: Commitment and extortion. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and MultiAgent Systems, AAMAS '07, ACM, New York, USA (2007)
39. Haselton, M.G., Buss, D.M.: Error management theory: a new perspective on biases in cross-sex mind reading. J. Pers. Soc. Psychol. **78**(1), 81–91 (2001)
40. Hauert, C., Traulsen, A., Brandt, H., Nowak, M.A., Sigmund, K.: Via freedom to coercion: the emergence of costly punishment. Science **316**, 1905–1907 (2007)
41. Heinze, C.: Modeling intention recognition for intelligent agent systems. Ph.D. thesis, The University of Melbourne, Australia (2003)
42. Herrmann, Benedikt, Thöni, Christian, Gächter, Simon: Antisocial Punishment Across Societies. Science **319**(5868), 1362–1367 (2008)
43. Hirshleifer, J.: Game-theoretic interpretations of commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 77–93. Russell Sage, New York (2001)
44. Hirshleiffer, J.: There are many evolutionary pathways to cooperation. J. Bioecon. **1**(1), 73–93 (1999)
45. Hofbauer, J., Sigmund, K.: Evolutionary Games and Population Dynamics. Cambridge University Press (1998)
46. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., Rizzolatti, G.: Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. PLoS Biology **3**(3), e79 (2005)
47. Imhof, L.A., Fudenberg, D., Nowak, M.A.: Evolutionary cycles of cooperation and defection. Proc. Nat. Acad. Sci. U S A **102**, 10797–10800 (2005)
48. Imhof, L.A., Fudenberg, D., Nowak, M.A.: Tit-for-tat or win-stay, lose-shift? J. Theor. Biol. **247**(3), 574–580 (2007)
49. Janssen, M.: Evolution of cooperation in a one-shot prisoner?s dilemma based on recognition of trustworthy and untrustworthy agents. J. Econ. Behav. Organ. **65**(3–4), 458–471 (2008)
50. Johnson, D.D.P., Fowler, J.H.: The evolution of overconfidence. Nature **477**(7364), 317–320 (2011)
51. Kautz, H., Allen, J.F.: Generalized plan recognition. In: Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI'1986), pp. 32–38. AAAI (1986)
52. Kraus, S.: Negotiation and cooperation in multi-agent environments. Artif. Intell. **94**(1–2), 79–98 (1997)
53. Leeds, Brett A.: Alliance reliability in times of war: explaining state decisions to violate treaties. Int. Organ. **57**(04), 801–827 (2003)
54. Lindgren, K., Nordahl, M.G.: Evolutionary dynamics of spatial games. Physica D: Nonlinear Phenom. **75**(1–3), 292–309 (1994)
55. Lockhart, Charles: Flexibility and commitment in international conflicts. Int. Stud. Quart. **22**(4), 545–568 (1978)
56. Macy, M.W., Flache, A.: Learning dynamics in social dilemmas. Proc. Nat. Acad. Sci. U S A **99**, 7229–7236 (2002)

57. Masuda, Naoki, Ohtsuki, Hisashi: A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. Bull. Math. Biol. **71**, 1818–1850 (2009)
58. Maynard-Smith, J.: Evolution and the Theory of Games. Cambridge University Press, Cambridge (1982)
59. Milinski, M., Semmann, D., Krambeck, H.J., Marotzke, J.: Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. Proc. Nat. Acad. Sci. U S A **103**, 3994–3998 (2006)
60. Miller, Geoffrey F., Todd, Peter M.: Mate choice turns cognitive. Trends Cogn. Sci. **2**(5), 190–198 (1998)
61. Nakahara, K., Miyashita, Y.: Understanding intentions: through the looking glass. Science **308**(5722), 644–645 (2005)
62. Nesse, R.M.: Evolution and the Capacity for Commitment. Russell Sage Foundation series on trust, Russell Sage (2001)
63. Nesse, Randolf M.: Natural selection and the capacity for subjective commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 1–44. Russell Sage, New York (2001)
64. Nowak, M.A.: Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press, Cambridge, MA (2006)
65. Nowak, M.A.: Five rules for the evolution of cooperation. Science **314**(5805), 1560 (2006). doi:10.1126/science.1133755
66. Nowak, M.A., Sigmund, K.: Tit for tat in heterogeneous populations. Nature **355**, 250–253 (1992)
67. Nowak, M.A., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner's dilemma. Nature **364**, 56–58 (1993)
68. Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity. Nature **437**,1291–1298 (2005)
69. Ohtsuki, H., Iwasa, Y.: The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. **239**(4), 435–444 (2006)
70. Ohtsuki, H., Iwasa, Y., Nowak, M.A.: Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature **457**(7601), 79–82 (2009)
71. Oyama, S.: Evolution's Eye: A Systems View of the Biology-Culture Divide. Duke University Press, Durham (2000)
72. Pacheco, J.M., Santos, F.C., Chalub, F.A.C.C.: Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput. Biol. **2**(12), e178 (2006)
73. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
74. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
75. Pereira, L.M., Han, T.A.: Evolution prospection. In: Proceedings of International Symposium on Intelligent Decision Technologies (KES-IDT'09), pp. 51–63. Springer Studies in Computational Intelligence 199, 2009
76. Pereira, L.M., Han, T.A.: Evolution prospection in decision making. Intell. Decis. Technol. **3**(3), 157–171 (2009)
77. Pereira, L.M., Han, T.A.: Intention recognition via causal Bayes networks plus plan generation. In: Progress in Artificial Intelligence, Proceedings of 14th Portuguese International Conference on Artificial Intelligence (EPIA'09), pages 138–149. Springer LNAI 5816, Oct 2009
78. Pereira, L.M., Han, T.A.: Intention recognition with evolution prospection and causal Bayesian networks. In: Computational Intelligence for Engineering Systems 3: Emergent Applications, pp. 1–33. Springer (2011)
79. Nichola, R., Aitken, D.: Uncertainty, rationality and cooperation in the context of climate change. Climatic Change **108**(1), 47–55 (2011)
80. Rand, D.G., Nowak, M.A.: The evolution of antisocial punishment in optional public goods games. Nature Commun. **2**:434 (2011)

81. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annu. Rev. Neurosci. **27**, 169–192 (2004)
82. Robson, A.: Efficiency in evolutionary games: Darwin, Nash, and the secret handshake. J. Theor. Biol. **144**(3), 379–396 (1990)
83. Roy, O.: Intentions and interactive transformations of decision problems. Synthese **169**(2), 335–349 (2009)
84. Roy, O.: Thinking before acting: intentions, logic, rational choice. Ph.D. thesis, ILLC Dissertation Series DS-2008-03, Amsterdam (2009)
85. Ruse, M.: Morality and commitment. In: Nesse, Randolf M. (ed.) Evolution and the Capacity for Commitment, pp. 221–236. Russell Sage, New York (2001)
86. Santos, F.C., Pacheco, J.M.: Risk of collective failure provides an escape from the tragedy of the commons. Proc. Nat. Acad. Sci. U S A **108**(26), 10421–10425 (2011)
87. Santos, F.C., Pacheco, J.M., Lenaerts, T.: Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proc. Nat. Acad. Sci. U S A **103**, 3490–3494 (2006)
88. Santos, F.C., Pacheco, J.M., Skyrms, B.: Co-evolution of pre-play signaling and cooperation. J. Theor. Biol. **274**(1), 30–35 (2011)
89. Santos, F.C., Pinheiro, F.L., Lenaerts, T., Pacheco, J.M.: The role of diversity in the evolution of cooperation. J. Theor. Biol. **299**, 88–96 (2012)
90. Santos, F.C., Santos, M.D., Pacheco, J.M.: Social diversity promotes the emergence of cooperation in public goods games. Nature **454**, 214–216 (2008)
91. Schelling, T.C.: The Strategy of Conflict. Oxford University Press, London (1990)
92. Searle, J.R.: The Construction of Social Reality. The Free Press, New York (1995)
93. Searle, J.R.: Making the Social World: The Structure of Human Civilization. Oxford University Press (2010)
94. Sigmund, K.: The Calculus of Selfishness. Princeton University Press (2010)
95. Sigmund, K., De Silva, H., Traulsen, A., Hauert, C.: Social learning promotes institutions for governing the commons. Nature **466**, 7308 (2010)
96. Skyrms, B.: Evolution of the Social Contract. Cambridge University Press (1996)
97. Skyrms, B.: The Stag Hunt and the Evolution of Social Structure. Cambridge University Press (2003)
98. Skyrms, B.: Signals: Evolution, Learning, and Information. Oxford University Press (2010)
99. Szabó, G., Fáth, G.: Evolutionary games on graphs. Phys. Rep. **446**(4–6), 97–216 (2007)
100. Tahboub, K.A.: Intelligent human-machine interaction based on dynamic Bayesian networks probabilistic intention recognition. J. Intell. Robot. Syst. **45**, 31–52 (January 2006)
101. Tomasello, M.: Origins of Human Communication. MIT Press, Cambridge (2008)
102. Traulsen, A., Nowak, M.A.: Evolution of cooperation by multilevel selection. Proc. Nat. Acad. Sci. U S A **103**(29), 10952 (2006)
103. Traulsen, A., Nowak, M.A., Pacheco, J.M.: Stochastic dynamics of invasion and fixation. Phys. Rev. E **74**, 11909 (2006)
104. Trivers, R.: Deceit and Self-Deception: Fooling Yourself the Better to Fool Others. Penguin Books, Limited (2011)
105. Trivers, R.L.: The evolution of reciprocal altruism. Q. Rev. Biol. **46**, 35–57 (1971)
106. van Hees, M., Roy, O.: Intentions and plans in decision and game theory. In: Verbeek, B. (ed.) Reasons and Intentions, pp. 207–226. Ashgate Publishers, Aldershot (2008)
107. Van Segbroeck, S., de Jong, S., Nowé, A., Santos, F.C., Lenaerts, T.: Learning to coordinate in complex networks. Adapt. Behav. **18**, 416–427 (2010)
108. Van Segbroeck, S., Pacheco, J.M., Lenaerts, T., Santos, F.C.: Emergence of fairness in repeated group interactions. Phys. Rev. Lett. **108**, 158104 (2012)
109. Vukov, J., Santos, F.C., Pacheco, J.M.: Incipient cognition solves the spatial reciprocity conundrum of cooperation. PLoS ONE **6**(3), e17939 (March 2011)
110. Wang, S., Szalay, M.S., Zhang, C., Csermely, P.: Learning and innovative elements of strategy adoption rules expand cooperative network topologies. PLoS ONE **3**(4), e1917, 04 2008

111. West, S.A., Griffin, A.A., Gardner, A.: Evolutionary explanations for cooperation. Curr. Biol. **17**, R661–R672 (2007)
112. Winikoff, M.: Implementing commitment-based interactions. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, pp. 868–875. ACM, New York, USA (2007)
113. Wooldridge, M., Jennings, N.R.: The cooperative problem-solving process. J. Logic Comput. **9**, 403–417 (1999)

Autonomous Mobile Robots as Technical Artifacts: A Discussion of Experimental Issues

Francesco Amigoni and Viola Schiaffonati

Abstract The definition of good experimental methodologies is a topic of growing interest in autonomous mobile robotics. Recently, researchers in this field have started to recognize that their experimental methodologies have not yet reached the level of maturity of other disciplines. In the effort of improving the quality of experimental activities, some proposals have been made to take inspiration from how experiments are performed in traditional sciences. However, a comprehensive analysis of the peculiar features involved in the experimentation of autonomous mobile robots intended as *engineering* artifacts is, to the best of our knowledge, still lacking. In this paper, we aim at contributing to fill this gap by discussing experiments in autonomous mobile robotics from an engineering point of view. We start by considering autonomous mobile robots as *technical artifacts*, namely as physical entities designed for a technical function and provided with a use plan. Then, we show that, due to the nature of the field, scientific and engineering aspects are strongly interrelated in the experimental activities performed to assess and evaluate autonomous mobile robots. To make our discussion more concrete, we refer to some examples taken from the specific robotic application of search and rescue of victims after a disaster.

F. Amigoni (✉) · V. Schiaffonati
Artificial Intelligence and Robotics Laboratory, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
e-mail: francesco.amigoni@polimi.it

V. Schiaffonati
e-mail: viola.schiaffonati@polimi.it

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_29, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

The emphasis on experiments and the effort in developing good experimental methodologies have gained growing attention in autonomous mobile robotics in the very last years. Basically, this community started to recognize that experimental methodologies have not yet reached the level of maturity of other disciplines and that objective and complete reporting of experimental results is not yet full part of the current research practice. To cope with this situation, a number of initiatives have been promoted, ranging from a workshop series on these topics [6], to special issues of journals [9], to European projects funded under different programs [5, 11, 13].

Against this scenario, in some previous papers [2, 4] we have analyzed how autonomous mobile robotics can benefit in taking inspiration from the way experiments are performed in traditional sciences. In this paper we widen our analysis by investigating experiments in autonomous mobile robotics from an *engineering* perspective. The starting point is showing that autonomous mobile robots can be considered as *technical artifacts*, in the sense of [17], namely as physical objects that have been built for fulfilling a technical function and have a use plan. Experimental assessment and evaluation of these particular technical artifacts involve a number of issues that, when properly elucidated, bring us to the conclusion that the engineering and scientific perspectives on experimental methodologies for autonomous mobile robotics are strongly coupled. To the best of our knowledge this is an original contribution, since we are not aware of any other work that addresses both the engineering and the scientific aspects of experiments in autonomous mobile robotics.

To better understand the rest of this work, we shortly introduce some basics of autonomous mobile robotics. *Autonomous mobile robotics* is an engineering field oriented to develop robotic systems that are autonomous in the sense that they have the "ability to maintain a sense of position and to navigate without human intervention" [14], in order to operate in places hardly accessible by humans or in cooperation with humans in common environments. Note that human operators are not completely excluded from autonomous mobile robotics, but they evolve from being active tele-controllers of the robotic systems to being more passive tele-supervisors of the same robotic systems. One of the application areas of major interest in autonomous mobile robotics, and the one we mainly consider in this work, is search and rescue [16], which develops robotic systems that are intended to assist human rescuers to detect and reach victims after a disaster. The tasks involved in these activities include, among others: moving between locations of disaster environments, building spatial representations (maps) of environments, and searching environments for victims.

This paper is organized as follows. After having summarized some results of our previous work (Sect. 2), we discuss how the notion of technical artifact, as analyzed in the current philosophy of technology, can enlighten some peculiar aspects of experiments in an engineering discipline, such as autonomous mobile robotics (Sect. 3). Then we focus on some experimental issues for technical artifacts, taking autonomous mobile robotics as a paradigmatic case (Sect. 4) and we conclude that, in this discussion, the engineering and the scientific perspectives should always be integrated (Sect. 5).

2 Toward Good Experimental Methodologies in Autonomous Mobile Robotics

In the last years, the interest in good experimental methodologies has increased within the autonomous mobile robotics community. In this section, we summarize some of our previous works, which discuss how the assessment of good experimental methodologies in this field can be inspired by some traditional experimental principles, and we reflect on how experiments in engineering disciplines present peculiar features with respect to experiments in the sciences. This section sets the general scenario in which the novel considerations introduced in this paper are discussed.

2.1 Taking Inspiration From Science

The autonomous mobile robotics community widely recognizes that its experimental methodologies have not yet reached the level of maturity of other disciplines. Despite the recognized importance of experiments for rigorously evaluating new approaches and for reporting them in an objective and complete manner, these approaches are not yet full part of the current research practice. In previous papers, we have discussed both how autonomous mobile robotics can be inspired by some traditional experimental principles in the process of assessing its experimental methodology [3, 4] and some epistemological issues that arise. Moreover, we have analyzed the experimental trends that emerge from the autonomous mobile robotics papers presented over the last 10 years at the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) [2].

Looking at the traditional experimental method, as shaped starting from the Scientific Revolution in the XVII century, it is possible to individuate some principles that constitute its core. They are comparison, reproducibility and repeatability, justification and explanation that, although do not exhaust the complexity of experimental method, represent some defining characteristics of experiments.

Comparison. The meaning of comparison in science is twofold. On the one hand, it means to know what has already been done within a field, to avoid the repetition of uninteresting experiments and to get hints on promising issues to tackle. On the other hand, it refers to the possibility for researchers to accurately compare new results with old ones. Comparison requires to embrace a sincerity principle so that anomalies and negative results, that can reveal something important, are reported.

Comparison in autonomous mobile robotics. The increasing use of public data sets (e.g., Radish [8], Rawseeds [11]) over which different systems can be run and compared is a clear sign of how comparison is acquiring crucial importance in autonomous mobile robotics. A recent emerging trend is toward the development of comparable implementations of systems, starting from their descriptions provided in papers and reports, also using the same code that was used in previous experiments (e.g., exploiting ROS [12] and OpenSLAM [10]). However, this trend is not strongly assessed yet, given that only slightly more than half of the 81 papers surveyed in [2] experimentally compare the proposed robotic systems with alternatives.

Reproducibility/repeatability. These principles are related to the idea that scientific results should undergo to the most severe criticisms in order to be confirmed. Reproducibility is the possibility to independently verify the results of a given experiment. Different experimenters must be able to achieve the same result, by starting from the same initial conditions, by using the same type of instruments, and by adopting the same experimental techniques. Repeatability concerns the fact that a single result is not sufficient to ensure the success of an experiment. A successful experiment must be the outcome of a number of trials, performed at different places and times in order to guarantee that the results have not been achieved by chance, but are systematic.

Reproducibility/repeatability in autonomous mobile robotics. The public distribution of code and/or problem instances (data sets) is a positive sign that experimentation within this field is moving toward a more rigorous approach. However, experiments involving several data sets referring to different settings (for example, to different kinds of environments, like indoor and outdoor ones) are still not so common. Hence, the implementation of similar experiments, that should draw the same conclusions to understand which parameters influence a robotic system, is very difficult. Moreover, the report of anomalies in performance to highlight which issues deserve further study in the future is rare. For example, only 5 papers out of the 81 analyzed in [2] report negative results, mainly related to situations in which the proposed robotic systems failed. On the positive side, we found in [2] that several papers use standard robotic platforms (e.g., commercially available ones) for experimental activities, easing the reproducibility of the results.

Justification/explanation. Justification concerns the drawing of justified conclusions on the basis of information collected during an experiment. It is not sufficient to collect as many precise data as possible, but it is necessary to look for an explanation, that is, all experimental data should be interpreted in order to derive the correct implications leading to the conclusions.

Justification/explanation in autonomous mobile robotics. From the survey in [2], it emerges that the weak attention given to statistical analysis of results hinders their justification and explanation. However, the use of several data sets promotes the derivation of well-justified conclusions. In particular, the correct behavior of robotic systems is usually verified according to ground truth (an optimal performance that represents a reference for evaluation) or to visual inspection. However, there exists a problem in generalizing the results obtained in an environment to other ones: how is it possible to rigorously demonstrate that a system works on instances for which ground truth is not available? We will consider again this issue in Sect. 4.

2.2 Experiments from Science to Engineering

As we have seen, traditional experimental principles can inspire the assessment of good experimental methodologies in autonomous mobile robotics and, indeed, in the very last years some works started to address in a more convincing way these principles. However, limiting the discussion on experimental methodologies in autonomous mobile robotics just to these scientific aspects is not sufficient, as autonomous mobile robotics is an engineering discipline, and experiments in engineering disciplines present peculiar features with respect to experiments in the sciences.

In general, when considering experiments in science and engineering, differences both at the level of the object and at the level of the purpose emerge. Not only do they focus on different objects (natural objects vs. human-made artifacts), but experiments have different purposes (to understand a natural phenomenon vs. to test an artifact). Most experimentation in engineering can be seen as a process composed of well-defined steps, namely, the definition of the goal, the choice of the factors to explore, the design and then the execution of the experiment, the analysis of data, and the drawing of the conclusions. For example, design tradeoffs have an impact on evaluation of experiments. Experimental methodology appears to be a list of strategies and well-organized resources that can be exploited whenever necessary and that can be collected under the label of *good experimental design practices*.

The purposes of experimentation in autonomous mobile robotics, however, cannot be exhausted by the activities labeled as experimental design practices. Its practice reflects the peculiar position of the discipline at the intersection of engineering and science. Robotic systems are human-made artifacts; accordingly, experiments have the goal of demonstrating that a given artifact is working with respect to a reference model (e.g., its requirements or its expected behavior) and, possibly, that it works better than other similar artifacts according to some metrics, thus making experiments closer to tests typical of engineering disciplines. At the same time, the most advanced robotic systems are extremely complex, and their behavior is hardly predictable, even by their own designers, especially when considering their interactions with the natural world, where complex environments are difficult, if not impossible, to model in a reasonable and manageable way. In this sense, experiments in autonomous mobile robotics have also the goal of understanding how these complex systems work and interact with the world and, therefore, are somehow similar to experiments in the natural sciences. We explicitly note that in this paper we are not considering the use of autonomous mobile robots to validate scientific (e.g., biological or ethological) hypotheses, like it happens for some bio-inspired robotic systems whose emergent behaviors are analyzed. As a consequence, the scientific aspects of autonomous mobile robotics we discuss are only relevant for the design and development of the robots themselves.

If experiments in autonomous mobile robotics from a scientific perspective have been discussed in the previous section, showing that an experimental methodology is slowly developing, in the following, we consider this experimental methodology from an engineering perspective. Here the notion of technical artifact seems to play a key role: engineering as an activity producing technology is a practice focused on the creation of artifacts and artifact-based services.

3 Technical Artifacts

Even if the discipline of autonomous mobile robotics deals with scientific tasks, such as to understand the behavior of robots in complex environments, nevertheless these robots are not naturally-occurring objects. In this section, we provide a definition of technical artifact and we contend that autonomous mobile robots are technical artifacts.

3.1 The Concept of Technical Artifact

In general *technical artifacts* are material objects that have been deliberately produced by humans in order to fulfill a practical function [17]. The term 'artifact' emphasizes the fact that these objects are not naturally-occurring. Technical artifacts present three key features:

- The *technical function* that is related to the question 'What is the technical artifact for?'
- The *physical composition* that is related to the question 'What does the technical artifact consist of?'
- The *instructions for use* that are related to the question 'How must the technical artifact be used?'

These three features are not independent of each other: to fulfill the technical function that the artifact is for, it has to be physically composed in a certain way and the user has to carry out certain actions, specified by the instructions for use. A technical artifact can be said to be a "physical object with a technical function and use plan designed and made by human beings" [17].

A technical artifact is the result of a *purposeful* human action; this seems to be a peculiar feature useful to differentiate technical artifacts from natural objects. Even if there is not any clear-cut distinction between the two and the dividing line is rather fuzzy, this does not mean that we cannot individuate some differences. These differences will play an important role in the following section to elucidate the different nature and goals of experiments in science and in engineering, as performed on natural objects and technical artifacts, respectively.

Natural objects are composed of physical objects and biological objects. Natural objects are characterized by properties that are described by physical, chemical, and biological laws, which are not human-made.

Let us compare first physical objects and technical artifacts. The latter are special physical objects with a function and a use plan, while general physical objects do not have any function. Think for example of an airplane and an electron. Even if an electron may perform functions in technological equipments, from a physical point of view this is irrelevant. Conversely, the function fulfilled by an airplane is an essential property of the object as a technical artifact. For instance, an airplane is built in order to fly by exploiting certain specific features assembled in a given way and in accordance with a use plan. It is worth noticing that, when ignoring the function and use plan, an airplane is a physical object in the sense that all of its relevant features and behaviors can be traced back to the laws of physics [15]. One may claim that also many of the phenomena studied by physicists today do not occur in nature, but are artificially produced in laboratories. Nevertheless, they remain natural phenomena studied and analyzed without any reference to their function and use plan.

Let us compare now biological objects and technical artifacts. At a first sight there is a clear difference between a bird and an airplane: the airplane has a function, while the bird has not or, at least, has a different type of function. Indeed, biological functions exist and this makes the distinction a little bit harder to establish. We argue that biological functions and technical functions are different on the following basis: while the former ones are usually related to the component parts of a biological object or to some behaviors of biological organisms, the latter ones are ascribed *both* to the parts of an artifact *and* to the artifact itself as a whole. For example, while the wings of a bird can be said to have the biological function of letting the bird fly, it is harder to say that the whole bird has flying as its biological function. On the other hand, the whole airplane can be said to have the technical function of flying. Moreover, and most importantly, the function of an artifact depends on use plans, while the biological functions do not: there is no use plan for the wings of a bird, whereas there are specific use plans for each of the parts of an airplane.

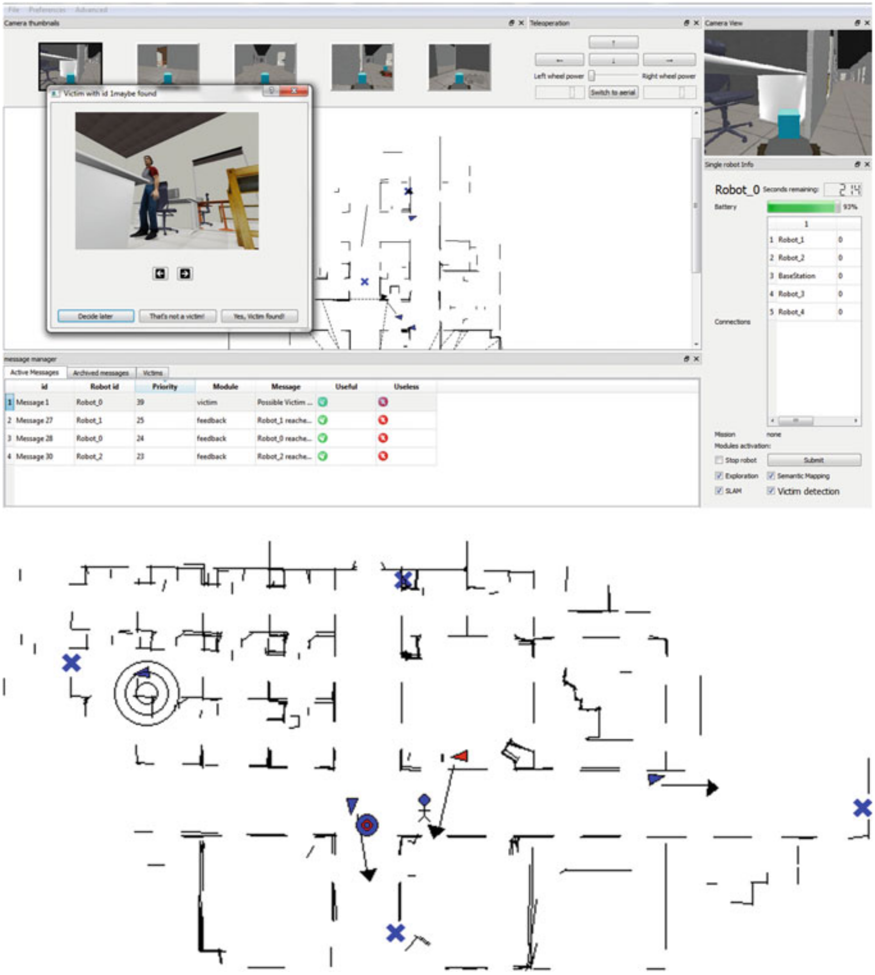
3.2 Autonomous Mobile Robots as Technical Artifacts

In this section, we argue that autonomous mobile robots are technical artifacts, according to the definition provided in the previous section, and we discuss some consequences of this fact. We focus on the search and rescue application in order to keep our discussion more grounded.

First, autonomous mobile robots for search and rescue are developed to fulfill a specific technical function, like exploring the largest possible amount of the area of an initially unknown environment in a given period of time, or exploring all the area of an initially unknown environment in the shortest possible time. During exploration, robots can be required to collect information about the presence of victims, about the possible paths that can be traversed by human rescuers, or about the structural stability of buildings. In general, the technical function can be defined by referring to the *task* the robots have to accomplish (e.g., looking for the largest number of victims within a given time) and to the *environment* in which the task is performed (e.g., indoor environment, like a collapsed building). Note that a task as defined before includes both an activity (e.g., search for victims) and a way to quantify performance in executing this activity (e.g., find the largest number of victims under some time constraints).

Being artifacts engineered for performing a technical function, the physical composition of autonomous mobile robots used for search and rescue is strongly related to that specific function. In particular, the physical components of robots are selected in order to cope with the intended task and environment. For example, using wheels for locomotion can be afforded only if the environment in which robots are expected to move is rather smooth, planar, and relatively empty. If this is not the case, locomotion needs to be based on legs or crawlers, or the use of aerial robots could be considered. Similarly, equipping a robot with a thermal camera could be useful if the task is to detect victims, while, if the task is to assess the structural stability of a building, other sensors could be more appropriate.

Finally, the use of autonomous mobile robots in search and rescue applications is a rather complex job and requires the presence of human operators to supervise operations and to actively intervene on the system in case of unexpected problems. Hence, the instructions for using these robots are significantly complex and usually require human operators to undergo some training. Commands that can be issued to the robots range from simple waypoint commands (i.e., “robot R, move to location A”, where A is in the known neighborhood of robot R) to more sophisticate commands (i.e., “robot R, explore areas along the North-West direction”). For example, Fig. 1 shows the graphical user interface of the PoAReT robotic system for search and rescue, which is composed of (simulated) autonomous mobile robots supervised by a human operator [1]. The bottom of Fig. 1 reports the different commands that the operator can issue to the robots: blue circles are waypoints, arrows are the preferred directions of exploration, and concentric circles are the preferred areas of exploration (blue crosses are the destination locations autonomously selected by robots themselves). Given the nature of search and



	id	Robot id	Priority	Module	Message	Useful	Useless
1	Message 1	Robot_0	39	victim	Possible Victim ...	✓	✕
2	Message 27	Robot_1	25	feedback	Robot_1 reache...	✓	✕
3	Message 28	Robot_0	24	feedback	Robot_0 reache...	✓	✕
4	Message 30	Robot_2	23	feedback	Robot_2 reache...	✓	✕

Robot_0 Seconds remaining: 214

Battery 93%

Connections

	1	
1	Robot_1	0
2	Robot_2	0
3	BaseStation	0
4	Robot_3	0
5	Robot_4	0

Mission none

Modules activation:

□ Stop robot Submit

☑ Exploration ☑ Semantic Mapping

☑ SLAM ☑ Victim detection

Fig. 1 The graphical user interface of the PoAReT system (*top*) and a detail of the commands that can be issued to robots (*bottom*)

rescue operations, timing in issuing commands and attention in guaranteeing the safety of all people (and, to a lesser degree, of all robots) involved are fundamental and should be part of the training of human operators.

4 Experimental Issues for Technical Artifacts

The way in which experiments for autonomous mobile robots are performed is deeply influenced by their nature of technical artifacts. Therefore, in this section we discuss some experimental issues of the discipline from an engineering perspective. In our discussion, we argue that, when dealing with some critical issues, scientific considerations re-emerge and have to be taken into account.

4.1 Technical Function and Experiments

The concept of technical function plays a central role in investigating the differences between natural objects and technical artifacts. Moreover, it is central in characterizing the different experimental approaches in science and engineering, keeping always in mind that there is not a clear-cut distinction between the two. As regarding experiments, the difference is not just related to the types of objects (natural objects vs. technical artifacts), but also deals with the function for which a technical artifact has been designed and built.

In engineering, experiments primarily evaluate technical artifacts according to whether and to what amount the function for which they have been built is fulfilled. For example, an autonomous mobile robot for search and rescue can be evaluated according to the amount of area it explores in a given period of time or according to the distance it travels (or the energy it consumes) to fully explore a given area. In science, experiments are primarily performed to discover new hypotheses that explain natural objects, to confirm or refute theories about them, and to choose among competing explanations. These natural objects can also be artificially produced by scientists, but they do not have any technical function and use plan and they are not evaluated according to them.

The focus on technical function and on the way it is evaluated in experiments calls for normative claims in the experimentation of technical artifacts. When making experiments on an autonomous mobile robot, for example, one of the first aspects to evaluate is whether the behavior of the robot is “good” or “bad”, and this normative evaluation is made with respect to a given reference function or sets of functions. More precisely, this evaluation is based on a *reference model* that has been used to develop the technical artifact, but also on some *metrics* that are used to measure the degree to which the artifact conforms to the model. For example, for an autonomous mobile robot employed in search and rescue operations, a reference model could prescribe that all victims in a given environment are found within 20 min, while a metric could measure the difference between the total number of victims in the environment and those actually found after 20 min.

The same normative evaluation does not apply to natural objects investigated in scientific experiments. An electron cannot be “good” or “bad”, although the explanation (hypothesis or theory) pertaining to electrons can be. As said, even if biological objects, as natural objects, have functions, they are not attributed to the biological object as a whole, but they pertain to parts of it, such as organs or behavioral patterns. Although it is true to say that a robot (as technical artifact) has a technical function, an animal living in the nature (as biological object) has not. It is thus impossible to make normative claims about a living organism as a whole. On the contrary, even if technical functions can be ascribed also to parts of technical artifacts, they can pertain to technical artifacts as wholes. Hence, technical artifacts, unlike natural objects, are interested by normative claims and this has an impact on the different ways experiments are conducted in engineering disciplines and in scientific ones.

The experimental evaluation of autonomous mobile robots as technical artifacts in terms of their technical function presents (at least) two critical aspects that involve not just technical issues, but also scientific ones. The first one is the gap between how the function is described in the project and how it is implemented in concrete. The second one deals with the difficulty of having a reference point to evaluate the system.

4.2 From Design to Implementation

Usually, the distance between the behavior that is planned (expected) at design-time and the actual behavior at run-time of an engineering artifact is not very large. Of course, the attempt to anticipate at design-time, through modeling, all (or most of) the possible problems that can arise at run-time is one of the fundamentals of engineering [18]. After all, the items considered at design-time, like needs, functional requirements, design specifications, and blueprints of the artifacts are globally intended to create good models of the final artifacts.

When considering autonomous mobile robotics, the above distance increases because the modeling at design-time of the interactions between autonomous robots and the real environments in which they are embedded becomes very complex. The reasons are the same that still hinder the development of reliable robotic simulators and that have been discussed in [3]. They can be summarized in the wide variability of situations in which autonomous mobile robots can find themselves and in the high difficulty in modeling their (largely unpredictable) interactions with the environment.

To cope with this difficulty, experiments performed in autonomous mobile robotics have two different targets, often only implicitly recognized by researchers: on the one hand, they are performed *to test* whether and how the design process has been correctly translated in practice; on the other hand, they are performed *to understand* the behavior of autonomous mobile robots when interacting with the world.

In the first case, an engineering perspective is adopted and autonomous mobile robots are experimentally tested to show that they work according to the technical function they have been designed for. In this case, a reference model and a metric to measure the closeness to this model are employed. For example, in the case of search and rescue, these tests include the ability of autonomously moving from a location A to a location B without colliding with any obstacle. The reference model is the correct execution of the task, namely location B is eventually reached and no obstacle is touched. The metric could measure the final distance between the actual position of the robot and location B and count the number of bumped obstacles.

In the second case, a more scientific perspective is adopted and autonomous mobile robots are subjects of experiments devoted to understand how these complex systems work when interacting with environments. For example, the PoAReT system [1] has been designed and built to represent environments using line segments and its behavior in non-nominal environments (i.e., those that cannot be naturally represented by linear features) cannot be easily predicted. Figure 2 shows the map built by the PoAReT system when working in an environment that contains highly non-linear features (in particular in the environment of Day 3 of the Virtual Robot Competition of the 2012 RoboCup Rescue Simulation League). From the analysis of the figure, it emerges that some features of the environment are not represented properly. However, predicting *ex ante*, before deploying the robotic system in the actual environment, which these features are and how incorrectly they are represented is almost impossible.

As another example, also related to search and rescue tasks, consider tests that are devoted to understand the behavior of a robotic system when it moves from a location A to a location B in presence of obstacles that were not explicitly modeled in the design of the robot. Indeed, a behavior for polygonal obstacle avoidance could be designed for the robot, but predicting its performance for round (non-polygonal) obstacles that can be encountered in real applications is very difficult. Elements that contribute to this situation include the errors in perceiving the obstacles (which are related to the design of the sensors mounted on the robots), the errors in deciding what to do to negotiate a perceived obstacle (which are related to the design of the control systems of the robots), and the errors in locomotion for avoiding the obstacles (which are related to the design of actuators mounted on robots). For example, a carefully-designed robot with all-shape

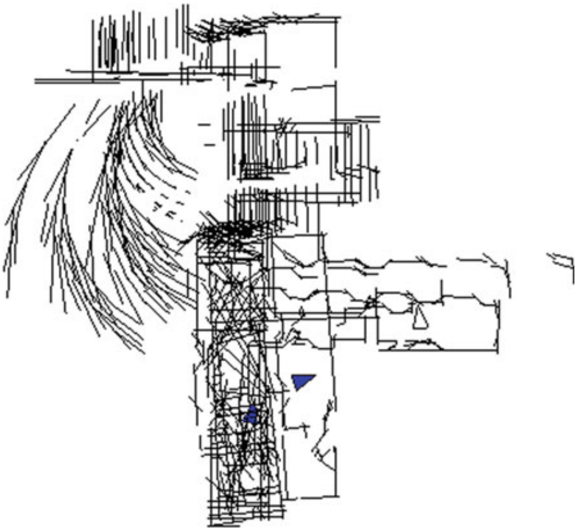


Fig. 2 A map built by the PoAReT system

obstacle avoidance behavior can still fail because some obstacles are not detected due to their transparencies, to the illumination of the environment, and to many other conditions that affect real environments. Moreover, another source of failure for the same carefully-designed robot could be a bug in the software program that decides where to move in order to avoid a detected obstacle. Finally, the very same robot can fail also because a properly-decided action is not performed as expected, for instance due to the surface of the environment that makes wheels slip.

Note that, while the second of the above causes (that related to control software programs) can be addressed using the classical software testing tools from software engineering, the first and third causes are specific to robotics and can be hardly addressed using tools mainly designed for software programs that interact with "controlled" environments, like those represented by users who can select only a limited number of options from menus. Accurately predicting and modeling all these aspects is beyond the current and near-future technical knowledge and, thus, they are not accounted for in the design of autonomous mobile robots.

4.3 Evaluation Without Ground Truth

The experimental evaluation of autonomous mobile robots for search and rescue is usually devoted to measure their behavior in some situations according to a number of metrics (and associated reference models) that can be divided in two classes. On the one hand, there are *non-functional metrics* (also called intrinsic), which are related to the efficiency of the internal behavior of the artifacts, like the amount of memory required to store a representation of the environment (like the map of a disaster site). On the other hand, there are *functional metrics* (also called extrinsic), which are related to the quality of the external behavior of the artifacts, like the time required to perform a given task (like moving along a straight line from location A to location B) or the length of a planned path between two locations.

While the non-functional metrics have mainly the role of helping in understanding the inner working of autonomous mobile robots while they perform some activities and, in this respect, they are supporting a more scientific approach to study the behavior of these artifacts, the functional metrics are mainly used to measure the degree of satisfaction of functions performed by autonomous mobile robots. Let us focus on this last issue.

As we have already discussed, usually an engineering artifact is experimentally evaluated against a reference model according to some metrics. For autonomous mobile robotics, the optimal, or reference performance, is often called *ground truth*. For instance, the ground truth for a map building task is the "optimal" (i.e., the most accurate) map of the environment, for example the blueprint. If the task is related to self-localization (i.e., to estimate robot location in a known map of the environment using only on-board sensors), the ground truth is the "real" position of the robot in the environment, measured in some way (e.g., by hand or by using

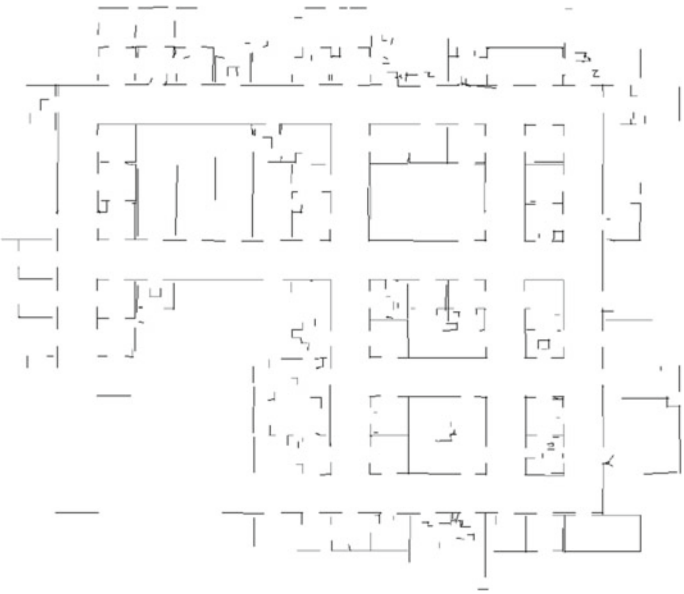


Fig. 3 Another map built by the PoAReT system

off-board sensors). As another example, when the task is searching an environment for victims, the corresponding ground truth includes the number and positions of victims actually present in this environment.

When the ground truth is available, comparing against it the performance of autonomous mobile robots is rather straightforward, given a metric. However, there are many situations in which the reference performance is not available. This constitutes a relevant difference between autonomous mobile robotics and many of the other engineering disciplines. Think, for example, of a map built by autonomous mobile robots that represent the internals of a damaged building. In this case, there is no way to get the "optimal" map of the damaged building with which the performance of the robots can be compared. For example, Fig. 3 shows a map built by the PoAReT system [1] in one of the test environments used in the Virtual Robot Competition within the 2012 RoboCup Rescue Simulation League. Without knowing the "optimal" map of this environment (i.e., the ground truth), it is very difficult to quantitatively assess the quality of the map built by the system, for instance to assess if the width of the corridors is correctly represented.

This opens a problem in the evaluation of the functions performed by these artifacts and also in their comparison. Indeed, without a reference model, robotic systems can be only compared against each other, lacking the possibility of an absolute comparison based on some idea of optimality.

5 Conclusions

In this paper we have discussed experiments in autonomous mobile robotics from an engineering point of view, showing that, due to the nature of the field, scientific and engineering aspects are strongly interrelated in experimental activities. We referred to the notion of technical artifact to elucidate the peculiar features of autonomous mobile robots intended as entities purposely created by engineers. In the experimental validation of these particular artifacts, some critical issues arise, like the gap between the designed and the actual behavior of autonomous mobile robots and the difficulty in obtaining a reference model for evaluating their performance, that make their experimental assessment closer to what happens in traditional sciences than to what happens in traditional engineering disciplines.

The main contribution of this paper can be summarized in the argument that considering only engineering aspects of experiments for autonomous mobile robots is not sufficient to cope with the complexity of the problems involved. Besides technical problems in the assessment of good experimental methodologies for autonomous mobile robotics, scientific and epistemological issues arise that influence the possible directions for future research. For example, to overcome the identified difficulty in evaluating the performance of a robotic system when the ground truth is not available, a pragmatic approach based on some strategies to assess the reliability of the results could be investigated, following what was proposed for the validation of simulation results in [3], under the influence of a fallibilist framework, such as that of [7].

Acknowledgments The authors gratefully thank the anonymous reviewers, Alberto Quattrini Li, the members of the PoAReT team (http://home.deib.polimi.it/amigoni/research/PoAReT.html), and the Fondazione Banca del Monte di Lombardia for the financial support.

References

1. Amigoni, F., Caltieri, A., Cipolleschi, R., Conconi, G., Giusto, M., Luperto, M., Mazuran, M.: PoAReT team description paper. In: RoboCup2012 Proceedings CD (2012)
2. Amigoni, F., Schiaffonati, V., Verdicchio, M.: An analysis of experimental trends in autonomous robotics papers. In: ICRA2012 (IEEE International Conference on Robotics and Automation) Workshop on the Conditions for Replicable Experiments and Performance Comparison in Robotics Research (2012)
3. Amigoni, F., Schiaffonati, V.: Good experimental methodologies and simulation in autonomous mobile robotics. In: Magnani, L., et al. (eds.) Model-Based Reasoning in Science & Technology, SCI 314, pp. 315–332 (2010)
4. Amigoni, F., Reggiani, M., Schiaffonati, V.: An insightful comparison between experiments in mobile robotics and in science. Auton. Robot. **27**, 313–325 (2009)
5. BRICS—Best practice in robotics, http://www.best-of-robotics.org/
6. EURON GEM Sig, http://www.heronrobots.com/EuronGEMSig/
7. Hacking, I.: Representing and Intervening. Cambridge University Press, Cambridge (1983)
8. Howard, A., Roy, N.: The robotics data set repository (Radish), http://radish.sourceforge.net/

9. Madhavan, R., Scrapper, C., Kleiner, A.: Special issue on characterizing mobile robot localization and mapping. Auton. Robot. **27**, 309–481 (2009)
10. OpenSLAM, http://openslam.org/
11. Rawseeds, http://rawseeds.elet.polimi.it/
12. ROS—Robot Operating System, http://www.ros.org/
13. RoSta—Robot standards and reference architectures, http://www.robot-standards.eu/
14. Siegwart, R., Nourbakhsh, I., Scaramuzza, D.: Introduction to Autonomous Mobile Robotics, 2nd edn. The MIT Press, Cambridge (2011)
15. Simon, H.A.: The Sciences of the Artificial, 3rd edn. The MIT Press, Cambridge (1996)
16. Tadokoro, S.: Rescue Robotics. Springer, New York (2010)
17. Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., Houkes, W.: A Philosophy of Technology. From Technical Artefacts to Sociotechnical Systems. Morgan and Claypool, San Rafael (2011)
18. Vincenti, W.G.: What Engineers Know and How They Know It. The Johns Hopkins University Press, Baltimore (1990)

Modelling Systems in Technology as Instrumental Systems

Maarten Franssen

Abstract Modelling is an extremely important aspect of the work of engineers. Ever since technology changed from a craft-based to a science-based practice, engineers have been engaged in modelling the artefacts they design, build and test. The modelling techniques they rely on, however, originate from the physical sciences. They work well for the technical devices, modelled as physical systems, that are the traditional products of the engineering disciplines. It is increasingly recognized, however, that modern technology consists in the implementation and operation of systems rather than single devices. The traditional conceptual framework of engineering, derived from the natural sciences, is ill-fit to model the hybridity and mereological complexity that are the key features of systems in technology. In this paper I present an approach to the modelling of systems in technology which is based on an incorporation of these two aspects from the start, represented in the notion of an instrumental system. I first show how the hybridity—the interaction between intentional action and causal processes—is taken care of in the basic structure of any instrumental system. Next I show how the representation of mereological complexity is taken care of through recursion. Finally relevance and potential applications of the approach are discussed.

1 Introduction: The Importance of Systems in Technology and How Engineers Model Them

Parallel to the industrialization of society in the course of the nineteenth century, there was a tendency for artefacts—entities designed, produced, implemented and operated in order to serve some purpose—to become larger and larger and to come

M. Franssen (✉)
Section of Philosophy, Delft University of Technology, Delft, The Netherlands
e-mail: m.p.m.franssen@tudelft.nl

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_30, © Springer-Verlag Berlin Heidelberg 2014

to consist of numerous components, many of which were only loosely coupled in a physical sense. It is this latter characteristic which makes us refer to such entities as systems rather than ‘things’ or ‘devices’. This colloquial identification of systems, which is equally at work in identifying systems in nature, such as ‘solar system’ and ‘ecosystem’, is not supported by the general definition of a system as an entity composed of interrelated elements: this definition does not make ‘systemhood’ dependent on the physical character of the interrelations, let alone their strength, and accordingly almost anything can count as a system. But this is not how the system concept is put to work. Artefacts are conceived as systems if they cannot be lifted up as one thing, with all of its components being lifted at once, by the mere forces that hold it together as one artefact, such as one could lift a steel furnace, a printing press or a rail engine off the ground. Still, as for example in the case of the GPS system, or its Russian competitor GLONASS, which consist of minimally twenty-four satellites orbiting the earth plus several ground stations and antennae, what is used is the entire systemic artefact, and to do so one must engage with the whole. For its operation it is necessary, but also sufficient, that many of its components are coupled only through the exchange of signals, typically electric or electromagnetic signals, involving an amount of energy almost negligible in comparison to the physical forces that make the wheels of a railway engine remain attached to their axles.

A particularly important subclass of such systemic artefacts are those artefacts where some of the components connected in this physically weak mode to each other or to other components are *people*. Notwithstanding the weakness of the interconnections from a purely physical point of view, for the functioning of the system these people are crucial, since they typically perform tasks that no device could perform, at least not equally satisfactorily or equally acceptably. Examples of such artefacts are factories, production firms, service-providing systems like electric power utilities and telecommunication networks, and rail, road and air transportation systems. Already in the late nineteenth century it was recognized that it was at the level of these systemic artefacts, with human components, that the success or failure of technological innovations such as the steel mill or electric light bulb would be determined [1], and that accordingly they should be designed just as methodically as the numerous technical devices that served as their components.

In the 1950s, due to the experiences gained in the military engineering projects during the Second World War, in particular the Los Alamos project, this recognition eventually culminated into the development of systems engineering as a separate branch of engineering. Systems engineering, however, focuses on the organizational part of designing and manufacturing artefacts: its systems are production systems, with the social processes—design teams, firm management, logistics—that this involves. Still, it was here that ideas within the conceptual framework of engineering on how to model large systems containing people as crucial components were first developed. Even though systems engineering contributed significantly to the successful execution of hugely complex engineering projects—space travel, ballistic missile technology—the conceptual contribution to a comprehensive grasp of systems depending crucially for their functioning on both technical devices and people, was modest. Systems engineering shared the conceptual outlook of all engineering: an exclusive orientation toward the natural sciences and their understanding of systems. Physical systems are understood to consist of multiple subsystems, themselves also understood as physical systems, which are connected by causal interactions: rigorous, law-like input–output connections, whether these connections are as strong as welds and rivets or as weak as electromagnetic waves. Human operators were modelled in the same way, to be interchangeable with causal feedback mechanisms, expected to match corresponding output to any input they received according to an exhaustive list of instructions. This 'physicalist' view of humans as components of complex engineered systems was soon recognized to be of little value for dealing with human–machine interaction as well as the interactions between humans. In particular it held no promise for understanding the many ways in which the behaviour of people deviates from what they are instructed to do or from what they have committed themselves to do.

With hindsight this physicalist or narrowly technical view was termed 'hard systems thinking', to be amended or even replaced by a new approach called 'soft systems thinking' [2].[1] The soft-systems approach, in contrast, focused entirely on the social relations between people and abstracted entirely from the physical processes in which these people are engaged. This matches the general approach to systems in the social sciences, for example in economics, where the representation of economic production is exhausted by production functions, i.e., by (conditional) decisions or intentions of the producers of goods. One of the few philosophers who has addressed the systemic character of the entities that both the natural and the social sciences deal with, Mario Bunge, explicitly defines social systems as consisting exclusively of humans, all (other) material objects being part of the environment of such systems [3]. As a result Bunge runs into difficulties when he wishes to introduce the notion of a technosystem, a system consisting partly of humans and partly of artefacts, as a special type of social system. For this inability to incorporate the material side of technology, 'soft systems thinking' was never accepted as a valid conceptual approach by the engineering community. Thus, during the period of about half a century since the birth of systems engineering, little was achieved because the people involved stuck to the ruling conceptual schemes, which offered only (purely) physical systems or (purely) social systems to work with.

To be sure, there are occasional contributions that suggest an awareness that more is necessary. Hubka and Eder [4] proposed that all technological action can be understood as the transformation of some object, and that all entities that realize such transformations have a common systemic structure. They coined the term

[1] One may assume that what was actually meant was 'hard-systems thinking' and 'soft-systems thinking', although the hyphen-less reading may be thought to fit the situation as well. But, deplorably, the hyphen is the most brutally ignored ingredient of English orthography in academic writing.

transformation system for this basic kind of system. The way Hubka and Eder further developed their proposal differs markedly from the modelling proposed here, however. From the start, they carve up a transformation system into subsystems that lump together the various types of components, and in this way isolated them from each other. In particular they distinguish as subsystems a *technical system*, a *human system*, a *management and goal system* and an *information system*. The mutual relations that can and must exist between these subsystems are not further analysed, nor is it at all clear how such further analysis might proceed. In particular no account is given of how the human subsystem refers partly to the 'driver' of the transformation, partly to the operators of the technical system, through which the transformation is realized, and possibly also partly to the operand, the object to be transformed. The technical subsystem contains all 'device' or 'hardware' elements of a transformation system, and only these elements. The information and goal subsystems, it would seem, should also be considered as to a large extent 'merged with' or 'contained in' the human subsystem. Notwithstanding their introduction of the notion of a transformation system, including their sketch of its basic structure, Hubka and Eder's interest concerned exclusively the—purely physically described—technical system.[2] A major aim of the modelling proposal described here, grounded in the notion of an instrumental system, is precisely to avoid the mutual isolation of the various types of components, which is explicit in Hubka and Eder but which underlies almost all of systems analysis in technology.

2 The Hybridity and Mereological Complexity of Systems in Technology; Instrumental Systems

To try to model systems in technology as being somehow composed of physical and social systems which are interlocked or meshed together seems a wrong approach altogether, because these two types of systems are articulated in quite different vocabularies: determinate causal interactions for physical systems versus intentional action resulting from deliberation and reasons for social systems. In technology, the intentional engagement of people with technical artefacts is central. It is what grounds the special character of systems in technology. Systems in technology, then, are essentially *hybrid*. They necessarily involve entities for the description of which we have developed quite distinct conceptualizations: on the one hand physical entities—technical devices, 'hardware'—and on the other hand social-intentional entities—people and associations of people. An artefact in complete isolation from the human world is in no relevant way different from a

[2] Note that their *Theory of technical systems* is a translation of the 1984 original with the German title *Theorie technischer Systeme*, which is again the second revised and extended edition of a book from 1974 entitled *Theorie der Maschinensystemen* ('Theory of machine systems').

natural object; it only 'plays out' its artefacthood in combination with a form of human engagement. The arrangements that this engagement leads to are systemic, because humans and artefacts play different roles in it. This is closely associated with a second basic characteristic of systems in technology: they are *mereological complex*. Not only can we distinguish between different entities playing different roles at the bottom level of technology, where a person engages with a thing for some purpose, but we can see this repeated once we look closer at any particular component, where we can again distinguish different subcomponents playing different roles. Almost all artefacts are themselves modular, consisting of components that have distinct roles or functions and whose physical make-up allows them to be connected and the various functions to support each other in realizing the overall function of the artefact. Therefore, artefacts, and artefact components in their turn, can be opened up, and complexes of entities, be they humans or devices, can be connected to, or introduced into, other complexes, leading to extended configurations, which improve existing functionality or realize new functions. This aspect has also been acknowledged by Hughes: "Systems nestle hierarchically like a Russian Easter egg into a pattern of systems and subsystems" ([5], p. 54).

In this paper I sketch a view of technological systems which puts both their hybridity and their mereological complexity up front. It represents the hybrid character of systems in technology by a hybrid type of model, in which the physical and the intentional or social aspects of technology are fundamentally merged from the start. It allows for the introduction of stratification into an arbitrary number of levels in a recursive way, emphasizing that what makes the complexity of technological systems difficult to understand and to handle is partly their sheer size and the number of their components and partly the fact that many of these components are people, performing a great variety of different roles.

I introduce the term '*instrumental system*' to indicate the comprehensive and basic type of system underlying technology.[3] Any instrumental system consists of three elements or components—a user, an instrument and an object—linked by two relations—a causal coupling of the instrument to the object and an intentional coupling realized by the user by manipulating the instrument-object complex in one way or another. The relevant intention in the latter relation is the intention to have the object undergo a particular transformation, a transformation that the instrument is capable of generating when linked in the right way to the object and when handled in the right way and circumstances by the user. An instrumental system, in short, is a system instantiated when an intentional agent handles a particular thing or device with the aim of changing the state of another thing. It is the sort of transformation that is the system's purpose that may serve to classify the system as a particular sort of instrumental system. For example, if someone—being the system's *user*—uses a nutcracker—being the system's *instrument*—to

[3] The first presentation of the notion of an instrumental system and of its stratified structure is [6]. I wish to acknowledge the important contributions of Bjørn Jespersen to the theory of instrumental system presented here.

crack a nut—being the intended transformation of the system's *object*—we have an instance of a *nutcracking system*.[4]

Formally we represent this as follows:

Instrumental system = $IT\langle User, CC\langle Instrument, Object\rangle\rangle$.

Here *CC* represents the *C*ausal *C*oupling relation instantiated when the instrument is applied in the right way to the object, and *IT* represents the *I*ntentional *T*ransformation relation instantiated when the user connects in the right way to the instrument-object pair. What is the 'right away' is determined by the nature of the instrument and the intention of the user. Not exclusively by the nature of the instrument, because any particular instrument—either a natural object or a technical artefact—can be used in various ways to contribute to the realization of various goals. In the next section I discuss several salient features of this conception of instrumental system.

3 The Main Characteristics of Instrumental Systems

First of all, it should be emphasized that 'user', 'instrument' and 'object' signify roles, not specific entities. In any concrete system, these roles are filled by specific entities. If I pick one of the fruits from a walnut tree and remove the nut, and then pick up a rock from the ground to crack the nut with, an instance of an instrumental system has come into being that can be called a *nutcracking system*. In this system, nut, rock and me fill the roles of object, instrument and user, respectively. It is not ambiguous to refer to the rock as the instrument of that system, since that what it is, in the predicative sense of 'is'. It gives rise, however, to an ambiguity in the meaning of the word 'instrument', since it can now be used to refer to the role and to the entity filling this role. This is an ambiguity that is difficult to avoid, and it is accepted in everyday language, where we can both say 'the president carries the launching codes' (referring to the role) and 'the president carries a blue briefcase' (referring to the person fulfilling the role). Context usually settles which of the two forms of reference is meant; this also applies to my usage in this paper. Similarly, when a nutcracker purchased in a shop is referred to as an instrument for nutcracking, irrespective of whether some nut is actually being cracked with it, the word 'instrument' is used not to refer to a role but to refer to a particular class of objects, viz. objects that have been designed and manufactured to fill the instrument role in a particular instrumental system. But it is perfectly possible that some instrument in the colloquial sense is never used as, and therefore never 'is', an instrument in the role sense at issue here.

[4] To be very precise, the intention in nutcracking is to crack a nutshell, rather than an entire nut. A cracked entire nut typically counts as a failure. Still, the instrumental system's object is the nut, not just the nutshell. If you could somehow isolate the shell from the nut to crack just the shell, you *ipso facto* have lost the need for cracking the shell.

Everyday language not only tends to blur the distinction between a role and its occupant, but also the distinction between the user role and the instrument role of an instrumental system. In a shop one would not so much ask for an instrument for nutcracking as for a *nutcracker*, whereas 'instrument for nutcracking' is an expression one would typically use if one was not sure whether there existed tools specifically designed for the cracking of nuts. Many instruments have names of the form 'φ-er', which indicates that what is named thus is, by design, an 'instrument for φ-ing'. Other examples are screwdriver, eraser, dishwasher, and so forth. In the conceptualization presented here, however, it is the *user* role of a φ-ing system, and by derivation any entity filling that role, which is indicated by φ-er. In this way, the conceptualization takes the active aspect expressed by the '-er' seriously: the tools we use do not really do the work for us, *we* do the work *through* our tools, and remain responsible for the outcome.

This blurring of the distinction between an instrumental system and its instrument happens particularly easy when the instruments are large-scale, complex systems rather than single devices. But even with 'systems' that are as complex as a national power grid, it is relevant to distinguish between the huge instrumental system of the kind 'public electric-power-providing system' and its instrument. To be a power-providing system, there has to be a user, often a national state, which sees it like that and brings it into existence as that kind of system and keeps it in existence by remaining actively engaged with it, for instance through paying the salaries and wages of its operators and securing an institutional context. We can conceive of the systemic instrument independently of its performing the instrument role in the power-providing system, and study its structure. Admittedly it is questionable whether such systemic instruments can remain intact for long in isolation, so to speak, since they will quickly lose their human components. But this depends on the larger environment. In a highly organized society a reserve army can easily be 'dormant' for ten or twenty years and then be brought into operational condition in a matter of hours. On the other hand, such systems may remain intact and operational for some time even when there is some—temporary or permanent—unclarity as to which agent or body occupies the user role. Compare the case of a driving system where the driver falls asleep for a few seconds or even minutes and then wakes up again and continues driving; it seems pedantic to insist that during those moments of sleep the driving system was inexistent because there was no-one intentionally operating it. Note also that as an instrumental system, a public electric-power-providing system can be operational without any customer actually tapping its power. It does not transform any particular customer from a state of being in the dark to having a lighted living room, since it is neutral with respect to what people will use their electric power for. Ideally, the system should survive everyone plugging in just as well as everyone unplugging. Rather, the system transforms prospective customers from a powerless state to a state of having permanent recourse to electric power, should they need it. Due to their extensive and diverse involvement of people, and the consequences this has for their implementation and operation, these systems are often called sociotechnical systems [7; 8, Chap. 5]. But what has been said about instrumental systems in general also applies here: the term ‘sociotechnical system’ is often used to refer to what is only the systemic instrument of a sociotechnical system in the sense of the conceptualization presented here, where it refers to a special kind of instrumental system.

These considerations serve to bring out two further important characteristics of instrumental systems. The first is the fact that they are dynamic, not static. If I obtain a walnut and pick up a rock to crack the nut with, not only do nut, rock and me fill the respective roles of an instrumental system, a nutcracking system, but they do so only as long as I hammer the nut with the rock. There was no nutcracking system there before I picked up the rock and there no longer is such a system once the nut has been cracked. The system exists in the cracking, so to speak. An instrumental system is conceived as a dynamic entity. However, it makes sense to distinguish different phases or states an instrumental system can be in: a system can be operational and can be idle. We have an operational system if the intended transformation proceeds as planned; an operational system is therefore truly a dynamic or process-like sort of entity. We have an idle system when everything is set to go and all the user needs to do is to ‘push the right button’ to set the system in motion and change its state to operational. This does not exhaust the possibilities, however. It can also happen that the instrument’s button is pushed, yet the system does not become operational and the object is not transformed as intended. If this happens, we have a malfunction occurring somewhere in the system.[5] The proposed modelling of instrumental systems right away suggests that there are various ways in which an instrumental system can malfunction. A particular nutcracking system can be malfunctioning, for instance, due to a malfunctioning instrument, as when we try to crack a nut with a nutcracker that has a broken hinge. This is the standard conception of malfunction. We can also have a malfunctioning nutcracking system, however, due to a malfunction in the coupling of the nut to the cracking instrument, causing the nut to be launched instead of cracked, or to a malfunction in the coupling of the user to the instrument, making the user’s hands slip off of the handles, for instance, because they are greasy. The instrumental-system conception of malfunction is therefore wider than the instrument or device conception of malfunction.

The second characteristic that is brought out is the fact that any instrumental system requires a perspective on what the system is supposed to achieve and how it is going to be achieved, in the sense of the particular actions that are required from the system’s user. This perspective resides in the user, by recognizing a way to use a particular entity for some purpose and then actually and intentionally using it so, in this way having this entity fill the instrument role of the corresponding instrumental system. The entity so used need not be a designed or engineered device, although typically it will be. But even if it is an engineered device, it need

[5] Malfunctioning systems can either be seen as included in the idle systems or as forming a system state of its own, next to operational systems and idle systems. I see no compelling reason to prefer the one to the other.

not be used as the kind of instrument it was designed as. One can use a nutcracker to crack a nut, but one could also use a rock, or a brick, or a monkey wrench to crack a nut, and one could use a nutcracker as a counterweight, or as a lever, or as a drumstick. It is the perspective of the user that determines the character of the instantiated instrumental system, although the freedom of the user is severely constrained by the laws of nature. No nutcracking system is likely to have a tuna sandwich figuring as its instrument.

The use of anything as an instrument starts with identifying two interfaces in it: one interface through which it is coupled to the object, the transformation of which is the user's aim, and another interface through which the user can manipulate the thing to realize the transformation. When I pick up a rock to crack a nut, I must both see how I will get a grip on the rock so that I can wield it and I must see where the rock can come down on the nut so that it will get cracked (rather than cut or pierced or squashed). Of course, with designed instruments, all this has been taken care of, if all is well, by their designers. Which is not to say that it is always entirely clear to a (prospective) user what the interfaces of a designed instrument are! But since instrumental systems are not dependent on the availability of designed instruments, in principle any tangible thing can fill the instrument role of an instrumental system: it is up to the prospective user of an instrumental-system-to-be to recognize two interfaces for some instrumental activity in a thing. Similarly, any tangible thing can fill the object role of an instrumental system. Even some of the planets of the solar system have served both roles, by helping to get space probes like the Voyager 1 and 2 into their intended trajectory through their gravitational pull, and by having received the marks of other probes that have landed on them, like the Viking 1 and 2 on Mars, on their surface. Less obvious perhaps, and therefore worthy of emphasis, is that 'any tangible thing' includes people. Many instrumental systems are aimed at the transformation of people, either physically or mentally, and we can use people, in particular people-as-bodies, as instruments just as much as we can use inanimate things.

No such freedom exists for what can fill the user role of an instrumental system. To be capable of defining an instrumental system through a perspective on what the system is supposed to achieve, the user of an instrumental system must be an agent, i.e., something capable of intentional states: expectations, aims, and the like. It need not be a human person, however. An association of people, in particular in the institutionalized form of a corporate agent—a business firm, organization, government or polity—can also be the user of an instrumental system, as long as it makes sense to ascribe to it certain aims and expectations. Under what conditions this makes sense, and how the ascribed aims and expectations are related to the aims and expectations of the individual people involved in it, is not a topic for further discussion in this paper, however. Nevertheless, it is a topic of crucial importance for instrumental systems that have a global scale, because their identification depends crucially on the possibility of identifying some instance as filling the role of its user, in 'whom' the perspective on what the system is and is about and how well it is functioning resides.

It may come as a surprise that the user of an instrumental system is modelled as an element or component of that same system. This may seem to be in conflict with the idea that the user is the user of something external to him or her. But the user is the user *of* an instrumental system only in the sense of being a component of it, belonging to it. She is not the user *of* the system in the sense that the system is what she is using. It is the instrument of the system that the user is using. User, instrument and object together form a system through the use that the user makes of the instrument in order to transform the object. Though the *fiat* that is any modeller's prerogative may do without reasons, in fact a decisive reason can be given for this 'inclusive' way of conceptualizing instrumental systems. A key concept in technology is 'function': what, in particular conditions that support this talk, things 'are for'. Technology, where we design and manufacture new kinds of things 'for' helping us in achieving our goals in life, could hardly do without the concept. There is no consensus concerning the exact meaning of the concept, however, and there are several more or less conflicting accounts or 'theories' of function around. The account that is accepted as the one that matches function talk in technology closest is the one proposed by Cummins [9]. Cummins' account makes the having of a function by an entity x dependent on x being a component of an enveloping system S. Then x has the function to φ if it is through φ-ing that x contributes to a capacity to ψ that the system S has. This works well for determining the function of the steering wheel of a car, since we can specify how turning the steering wheel makes it possible that the car as a whole can be driven by its driver. The account works similarly in the biological realm for determining the functions of the organs of organisms. However, difficulties arise if we wish to ascribe a function to a complete technical artefact, which requires only the handling of a user to come to do what it was designed and made for. Cummins's account of function can only be saved here if we can point out a system of which the complete artefact is a component and which has a capacity partly due to the artefact's operation. My contention is that it is precisely the instrumental system made up by artefact, user and object that is this system. It is through the role that instruments play in instrumental systems that we can meaningfully say, using Cummins' functional analysis, that instruments have functions.[6]

4 Taking Care of Mereological Complexity: Stratification of Instrumental Systems into Levels

The introduction of the notion of an instrumental system is merely the first step toward the modelling of the full complexity that systems in technology can have. A crucial second step is the introduction of stratification. This is done by way of

[6] This cannot work in biology, but there we may acquiesce in finding it impossible to ascribe a function to an entire organism. On the other hand, organisms could be seen as having functions derived from the role they play in ecosystems.

recursion. Not only can, in this way, an arbitrary number of levels be given to any instrumental system, but the hybrid character of the user-instrument pair can be repeated at every level as well. Each of the roles of user, instrument and object is in principle open to be filled by an entity that is itself systemic in character in conceptually the same way, that is, analysable as consisting of sub-entities that fill at least two of the roles of user, instrument and object, linked by specific relations. However, it will become clear below that for each of the three roles there are restrictions on what further structure the entities that fill them can have.

A relatively simple example showing this layered structure is a taxi. In a taxi, the instrument utilized by the user to be transported is analysable as consisting of an instrument role—filled by the car—and a user role—filled by the driver of the car. The taxi driver is related to the car in a way that differs from the way a person driving a car to transport herself or other people or objects is related to the car. Such a person uses the car as the instrument of an instrumental system we can refer to as a simple driving system. To actually perform that role, one must intend these transformations and operate the car accordingly. The taxi driver, in contrast, does not intend the transportation of either himself or even his passengers; what he intends is to serve his clients so as to receive his payment. Typically, this will involve the transportation of the client, but ultimately it is the client who will decide this. Suppose that a taxi is ordered to a home address, where the client leaves the house, gets into the car and tells the driver to start driving, say, to the airport, but next, after the first curb, tells the driver to stop and let him out. The client did not intend to be transported at all but intended to be seen driving away in a taxi. The relation that the taxi driver has to the car is therefore of a sort different from the relation binding the user of a simple driving system to the car-cargo complex: it is partly active and intentional—the taxi driver performs his role of his own free will, we may assume, and is conscious of what he is doing in manipulating the car and why these precise manipulations are required—but also partly passive—he operates his car in response to the instructions of his customer, and he is generally not conscious of why he is receiving these instructions, or at least it is not relevant for his ability to manipulate the car adequately to be conscious of this. Up to certain limits, that is. Up to certain limits of what he can and will be prepared to do in response to instructions he receives, limits set partly by considerations for his own safety and partly by the laws of the society that he is a member of. And up to certain limits to what he must understand about the reasons for being instructed as he is in order to do his job properly, limits set by character of the system that he is part of.

Taking up the conceptualization introduced in the previous section, any particular kind of instrumental system, say a simple driving system, can be represented by giving the user, instrument and object roles in the basic model names proper to the type of system formed:

$$\text{Simple driving system} = IT\langle Driver, CC\langle Vehicle, Passenger/cargo\rangle\rangle.$$

From the word ‘vehicle’ one cannot derive that ‘Vehicle’ here refers to an instrument role. However, that ‘Vehicle’ indicates an instrument role is unambiguously fixed by its position in the formal scheme. It is the syntax of the modelling language that decides this.

An assisted driving system can then be represented as follows:

Assisted-driving system = $IT\langle User, CC\langle IC\langle Driver, Vehicle\rangle, Passenger/cargo\rangle\rangle$.

We normally call the person who controls the pedals and steering wheel of a car the driver. The user role of the entire assisted-driver system is now no longer appropriately designated as ‘driver’. Given that we adopted the term ‘assisted-driving system’ for this kind of system, we could refer to this particular user role as ‘assisted driver’, but for clarity’s sake I will stick to ‘user’. An assisted-driving system is a special kind of instrumental system in which the instrument role is filled by an entity that itself has a structure in which two of the three basic roles of an instrumental system occur, linked, however, by a new relation *IC*, for Intentional Coupling:

Systemic instrument: $IC\langle User, Instrument\rangle$.

If we accept that any instrumental system can be indicated as a φ-ing system, where φ-ing is what the user does when he or she uses the system’s instrument to φ the system’s object, then an assisted driving system is a form of instrumental system that can be termed an assisted-φ-ing system:

Assisted-φ-ing system = $IT\langle User, CC\langle IC\langle User', Instrument\text{-}for\text{-}\varphi\text{-}ing\rangle, Object\rangle\rangle$.

The prime next to the second *User* is there to indicate that the two roles *User* and *User′* are different roles, which can be filled independently of each other, though both roles are of the kind ‘user’, i.e., subject to the restrictions governing all entities that fill this kind of role. It needs some argumentation that, for example, the role of a taxi driver can count as a user role with respect to the vehicle as instrument. The primary reason for doing so is that the person filling this role sets the vehicle in motion by actively and intentionally handling it, similarly to the way the user of a complete instrumental system sets the system in motion. What makes a taxi driver different from a driver corresponding to the user role of a full instrumental system is the relation connecting the person to the instrument: the relation *IC*, through which the user of the systemic instrument $IC\langle User, Instrument\rangle$ is connected to *its* instrument, is different from the *IT* relation connecting the user of an instrumental system to the $CC\langle Instrument, Object\rangle$ complex. The handling of the car by the taxi driver is not motivated by an intention to *transform*; there is not even necessarily an object to be transformed, and the characterization of the systemic instrument does not require an object role. The driver of a taxi can be just that but not transform anything, e.g., when he is ordered to stay put, or to drive around just to soothe a baby to sleep, or in the deceit story above. In the first case, we have an idle assisted-driving system, in the second case an assisted-driving system but not an assisted-driving-to-transport system, in the third case we do not even have an assisted-driving system but rather a deceiving system. In none of these cases, however, does the taxi driver control the nature of the system; his role is to follow instructions, and the instructions may reveal to him, if perhaps only gradually or partially, what the nature of the overall system is in which he, jointly with his vehicle, figures as an instrument. That is why he is a component of the system in the instrument slot. But since his following of instructions is an *intentional* activity, as is his translation of the instructions into the operation of the vehicle, in which he enjoys considerable freedom, making his hands-on manipulation of the car indistinguishable from the manipulation of the user/driver of a simple driving system, the kind of rule he fills in the system is sufficiently similar to the role of end-user to classify both as being of the user type.

Next to a systemic component consisting of a user and an instrument role, we can also have a systemic component consisting of an instrument and an object role:

Systemic instrument or object: $CT\langle \mathit{Instrument}, \mathit{Object}\rangle$.

Here *CT* refers to a *C*ausal *T*ransformation of an object, which takes place entirely by physical forces, without this requiring a form of intentional button-pushing by a user. To analyse a component of an instrumental system as having this structure typically is in order when parts of the system are automated. For example, designed devices are typically modular, consisting of subunits responsible for subfunctions, with the way the various subunits are hooked up realizing the overall function of the device. Another example is where an object needs to undergo a transformation first to 'prepare' it for the transformation process that a user inflicts upon it by using some instrument on it. Take a driving system where the object to be driven, and thereby transported, is a cat. If a person has to be driven to his or her destination it suffices to make that person sit on one of the car seats, but it is not wise to adopt the same approach with a cat. The cat can be transported only if it is first put in a cage; otherwise there is a considerable danger that the cat is going to interfere with the driving process, by clawing its nails into the driver's neck or by ending up under the brake pedal. This could be analysed as $CT\langle \mathit{Cage}, \mathit{Cat\text{-}cargo}\rangle$ (taking into account that 'cat' is hardly a role that a particular animal can fill, but 'cat-cargo' is). However, this may not be considered entirely satisfactory, in view of the fact that an entity with the structure $CT\langle \mathit{Instrument}, \mathit{Object}\rangle$ is a dynamic, process-like entity, due to the transformation aspect of the *CT* relation, and it could be objected that what is transported on the back seat of the car is an object, not a process. This can be remedied by viewing the cat-cage complex as an instance of $CC\langle \mathit{Instrument}, \mathit{Object}\rangle$. This, however, is not entirely satisfactory either, since there is something unfinished, preparatory, about the static $CC\langle \mathit{Instrument}, \mathit{Object}\rangle$ complex, matching its status as a subsystem in an instrumental system, waiting to be 'started off' by that system's user. To deal with this case, therefore, it makes sense to introduce a further element into the analytical framework: the syntactic operator *APP*(.), which takes as its input either an entire instrumental system or the 'reduced' system $CT\langle \mathit{Instrument}, \mathit{Object}\rangle$. So,

Output of a φ-ing system $= \mathit{APP}(\mathit{IT}\langle \mathit{User}, \mathit{CC}\langle \mathit{Instrument\text{-}for\text{-}}\varphi\mathit{\text{-}ing}, \mathit{Object}\rangle\rangle)$

indicates a ‘φ-ed object’, an object which has been φ-ed intentionally by being put through the act of φ-ing by a user applying an instrument-for-φ-ing to it. The advantage of adding this descriptive element to the conceptual framework is that it allows us to explicitly introduce agenthood with respect to subfunctions and subtasks into the description of a larger system. If a cat is to be transported by car, then *someone* must first put the cat in a cage, and accordingly someone has a *responsibility* to see to it that the cat gets into the cage. Only if subfunctions are completely automated, typically within an complex engineered instrument, will this responsibility have dropped out and can we do with *APP*(*CT*⟨*Instrument-for-φ-ing*, *Object*⟩).

Each of the three main roles of an instrumental system can be analysed as having the structure APP(*IT*⟨*User*, *CC*⟨*Instrument-for-φ-ing*, *Object*⟩⟩). The caged cat on the back seat of a car, thus prepared for transportation, is an example of an *object* with that structure. For an *instrument* with that structure, take the lock on the door of a car, or a room. The car is a instrument for driving, but typically, to fill that role it first needs to be unlocked. So the car filling the instrument role of a driving system can be analysed as having the structure *APP*(*IT*⟨*Unlocker*, *CC*⟨*Car key*, *Vehicle*⟩⟩). Usually, it is the driver who fills the ‘unlocker’ role, and making the act of unlocking the door of the car explicit would not easily occur to anyone as being of much use. However, there are many situations where responsibilities are so distributed that making such roles explicit in the modelling is crucial for understanding how the system works and how it could fail to work. The same applies to *user* roles. The user of a simple driving system, for instance, needs to have sufficiently sharp vision in order to be able to fill this role adequately. If the person who is going to fill that role fails to have sharp vision ‘by nature’, this can be remedied by having him or her put on a pair of glasses, making the entity filling the driver role analysable as having the structure *APP*(*IT*⟨*Fitter*, *CC*⟨*Glasses*, *Driver*⟩⟩). In a simple driving system, the role of fitter, the ‘locus’ of the responsibility for seeing to it that the driver of the system is fit with a pair of glasses, will typically be filled by the same person who does the driving, and it may seem conceptual overkill to make it explicit. But here as well, it may be entirely relevant for understanding how the system works and anticipating how it may fail to work, to track responsibilities. If a simple driving system is modelled as *IT*⟨*APP*(*IT*′⟨*Fitter*, *CC*⟨*Glasses*, *Driver*⟩⟩), *CC*⟨*Vehicle*, *Passenger/cargo*⟩⟩, then this points out a recognition by the end-user of this system, whose perspective is represented in the definition of the system, that fitting the driver with glasses is required for the system’s well-functioning, and that accordingly someone, the ‘fitter’, has the responsibility of making this happen. These considerations more obviously apply to any system that has people filling user roles somewhere within the system’s complex instrument, where it seems highly relevant to make explicit in the modelling whether there are particular conditions that these people must satisfy for filling a role, and these conditions typically become satisfied by having them undergo the sorts of transformations that instrumental systems generally realize: wearing protective clothing, being vaccinated, being properly trained, and so forth.

The approach sketched so far, to represent components of an instrumental system as themselves structured in a similar way, is a choice for a maximally parsimonious conceptualization, which needs just three different kinds of roles—'user', 'instrument' and 'object'—, four different kinds of relations—*IT*, *CC*, *IC* and *CT*—and an *APP*(.) operator as building blocks for representing arbitrary levels of complexity. What was said in the previous section on the requirements for an entity to fill any of the three roles in the basic instrumental system applies equally to these roles when they appear at lower levels, that is, *within* an entity filling any of these roles, when we open it up for further analysis and reveal its structure in the same terms.

There are restrictions to the type of systemic analysis that the components of instrumental system allow. No role in an instrumental system can, for example, be filled by an entity having the following structure:

*Systemic instrument or object: $XX\langle User, Object\rangle$.

Here *XX* stands for any of the four relations occurring in instrumental systems: *IT*, *CC*, *IC* or *CT*. Within the setting of instrumental systems, a user can be related to an object, bringing about its transformation, only through the mediation by an instrument. There may be situations where this seems to be violated, in particular when a person uses his or her body as an instrument, for example, when a parent throws his or her body in the trajectory of a bullet fired at a child, in order to prevent the child being hit by the bullet. In such a case, the person fills the user role and also the instrument role. What is transformed is either the bullet, from continuing its trajectory to being stopped in its trajectory, or the child, from being targeted by the bullet to being delivered from it. In a case like this, we may prefer to say that the person's *body* fills the instrument role. However, a person filling a particular user role is always an embodied person, since as a user the person must physically manipulate an instrument. Even in the case of an assisted-driving system, say a taxi, giving instructions to the taxi driver is always some physical process.[7] In using an instrument, a person must identify an interface in the entity serving as instrument for coupling it to the object and a second interface for coupling to as a user. When someone uses his or her body as an instrument, these interfaces are in a sense internal to the person. A person is naturally geared to his or her body. There are limits, however, to being allowed to say that someone uses his or her body to achieve something physical. If a small child needs to be transported by car and instead of 'caging' it, by strapping it for instance, the child is kept in check by entertaining it with games and conversation, no physical instrument plays a role at all: the person doing the entertaining does not use anything in the instrumental-system sense. There is a difference between using your body to catch a bullet meant for your child, or even using your fingers to remove a hair from your diner jacket, and going through the automated routines of talking, breathing, staying upright, and

[7] Metaphysically, this may not be so straightforward: the 'is' in this statement is arguably not the 'is' of identity but rather the 'is' of realization or constitution.

the like. You cannot say that you use your lips, teeth and tongue to talk to the child; this would come dangerously close to saying that you use your lungs and your pectoral muscles to breath. Which is what we definitely should not say. There is no effort in expanding your chest while breathing or pursing your lips while speaking. There *is* effort in picking up a hair from your diner jacket, and certainly in throwing your body in the trajectory of a bullet.

Apart from particular structures that no entity filling a role can have, like $\mathit{XX}\langle\mathit{User}, \mathit{Object}\rangle$, there are constraints on the structure that an entity filling a user role can have. Arguably, nothing that has the dynamic, process-like character of an instrumental system or of its ‘incomplete’ user-instrument and instrument-object forms can at the same time be an entity capable of intentionality as required for a user role. Consequently, an entity filling a user role is never open to analysis that gives it this sort of structure. The only further structure that seems acceptable for the user role, at whatever level it occurs, is the $\mathit{APP}(.)$ kind, meaning that a person or person-like entity filling a user slot is first modified by ‘passing through’ an instrumental system.

In contrast, both the instrument role and the object role of an instrumental system, at any level where they occur, can be filled by an entity that itself has the structure of a complete instrumental system $\mathit{IT}\langle\mathit{User}, \mathit{CC}\langle\mathit{Instrument}, \mathit{Object}\rangle\rangle$. There are two general types of systems that have a structure involving this. The first has an *instrument* that is a complete instrumental system:

Benefiting-from-φ-ing system $= \mathit{IT}\langle\mathit{User},\ \mathit{CC}\langle\underline{\mathit{IT}'}\langle\underline{\mathit{User}'},\ \underline{\mathit{CC}'}\langle\underline{\mathit{Instrument\text{-}for\text{-}}\varphi\mathit{\text{-}ing}},\ \underline{\mathit{Object}'}\rangle\rangle,\ \mathit{Object}\rangle\rangle.$

Here the user of the benefiting system uses an instrumental system to change the state of an object from being ‘without benefit’ to having benefited from the activity of φ-ing. Take a person who drives a taxi to earn an income, and assume the person is independent, that is, not employed by a company. We then have an instance of a *Benefiting-from-assisted-driving system*:

$$\mathit{IT}\langle\mathit{User},\ \mathit{CC}\langle\mathit{IT}'\langle\mathit{Client},\ \mathit{CC}'\langle\mathit{IC}\langle\mathit{Driver},\ \mathit{Vehicle}\rangle,\ \mathit{Passenger/cargo}\rangle\rangle,\ \mathit{Object}\rangle\rangle.$$

Typically, one and the same person, the ‘taxi driver’, fills three different roles: first the system’s user, that is, the person whose perspective and corresponding actions define the benefiting system and render it operative, i.e., cause someone to benefit; second the driver role; and third the object role, since it is the person driving the taxi who benefits by earning an income. The three roles could, however, be filled by different persons: someone could work as a taxi driver in order to earn money exclusively for some friend or relative, or for some charity, and a taxi driver’s relative could fill in (‘drive in’) for him in case of illness and hand him the income, because he urgently needs it. And these two cases could even be combined, causing the system user, driver and benefiting person roles to be filled, at least momentarily, by three different individuals.

The second general type of system has an object that is a complete instrumental system:

φ-ing-system-regulating system $= \mathit{IT}\langle \mathit{User}, \mathit{CC}\langle \mathit{Instrument}, \underline{\mathit{IT'}}\langle \underline{\mathit{User'}}, \underline{\mathit{CC'}}\langle \underline{\mathit{Instrument\text{-}for\text{-}\varphi\text{-}ing}}, \underline{\mathit{Object}}\rangle\rangle\rangle\rangle$.

Here the user of the regulating system uses an instrument to transform some φ-ing system. As an example, take a civil authority that regulates the behaviour of driving systems at a particular location—people driving a car over some local road—by placing some instrument on or near the road. This leads to a *Driving-system-regulating system*:

$\mathit{IT}\langle \mathit{User}, \mathit{CC}\langle \mathit{Instrument}, \mathit{IT'}\langle \mathit{Driver}, \mathit{CC'}\langle \mathit{Vehicle}, \mathit{Passenger/cargo}\rangle\rangle\rangle\rangle$.

(Some of these driving-systems could actually be assisted-driving systems—taxis and freight lorries; this complication is ignored). The instrument could be anything that makes the drivers adapt their driving behaviour: a speed bump, a traffic light, a road sign, an automatic speed-triggered camera, or a police officer with a notebook. Why any of these instruments can be expected to influence driving behaviour depends, of course, on the wider context. Generally, in calculating the expected effects of any regulating instrument, knowledge about the system to be regulated is crucial; in this case it is largely exhausted by knowledge about the mental states of car drivers. These states reflect not just the presence of an institutional setting intended to punish the trespassers of traffic rules, but also drivers' estimates of and experiences with the extent to which this setting succeeds in actually punishing them.

5 Relevance and Applications

The explications given in the previous section for the modelling of particular types of instrumental systems through an analysis of how entities filling a user, instrument or object role contain in their turn further entities that again fill user, instrument and/or object roles, will already have revealed some of the motivation underlying the proposal for modelling systems in technology presented in this paper. Together they serve to emphasize that in technological systems complexity and hybridity typically go hand in hand and reinforce each other. Much of the complexity of such systems is a form where the hybridity underlying the system character of any technical system—the intentional initiation across an instrument's interface of a causal process aimed at the realization of a desired outcome—is repeated at deeper levels where such intentional couplings occur at, relatively, smaller scales. There is a clear difference between someone using a car to drive to work and using a taxi to get to work. The first system can be understood as containing just one layer, the basic one of user-instrument-object, the second requires two layers, as there are two user-initiated manipulations at work: one of the car by its driver and one of the driver by the client.

To be sure, increased technical complexity does not necessarily go hand in hand with increased hybridity. Modern technical devices, for example a modern airliner like the Airbus 380 or the Boeing 757, are more complex in having more technical layers than older and simpler artefacts. Some researchers who are interested in understanding and dealing with the complexity of modern technology, for which they coin the notion of *engineering systems*, use the number of layers into which the elements of such systems can be decomposed—components consisting of components consisting of components, and so forth—as a defining criterion, and reserve no special role for some of these components being human beings [10, Chap. 2]. However, notwithstanding the indeed stunning complexity of modern technical devices in this sense, which requires the engagement of many different areas of technical expertise in their design, manufacture, operation and maintenance—another indicator of device complexity for the supporters of engineering systems—still the presence of humans adds incredibly more complexity to any technical system. Compare, for example, any modern aircraft to the civil air transport system in which such aircraft figure as components, but which contains additionally the aircraft crew, the passengers, the air traffic controllers distributed all over the globe, the personnel of airlines and airports, the airline and airports executives and owners, and so forth. For aircraft, any of its technical components, however minute, may have a decisive impact on its behaviour. This carries over to the functioning of the civil air transport system, extended to all other devices that are part of it, but to this is added the decisive impact that any of the human components can have.

A small example serves to show that the human involvement in a system can easily surpass the system's technical complexity. Take a situation where a municipality has installed a traffic control system regulating lane access through overhead signs, which are operated by a traffic controller, and let the traffic controlled consist of a single taxi. We have just two devices—overhead signs and vehicle—but four persons—municipality, traffic controller, taxi driver, client—and what happens will not just depend on whether the signs come on when the operator pushes the corresponding button and whether the car's engine will continue to run, the car's tires hold out and the petrol tank remains sufficiently filled, but also on whether the operator has received clear instructions on when to limit or extend lane access, whether the taxi's client is in a hurry and will urge the driver on, whether the taxi driver is hard on money and inclined to comply, and so forth.

This example also serves to emphasize that from the instrumental-systems point of view, a particular physical situation or entity never *is* a system in the sense that it is *one particular* system. To every person present corresponds a perspective defining a system. In the above example, next to the driving-system-regulating system, we can distinguish an assisted-driving system, a benefiting-from-operating system, and a benefiting-from-assisted-driving system. We need to take into account all of these to fully understand why the entire situation 'works' for all participants, what is required to make it work and what can cause its working to be disrupted. If it did not work for the taxi driver, for instance, meaning that the taxi driver failed to benefit from it by earning a living, then neither would it work for the client, since the client will not be transported. And if the situation failed as a regulating system, say, because the instructions to the operator are ambiguous or conflicting, then it might also fail as an (assisted-) driving system, since no access to a driving lane might be given at all.

An important contribution of this way of modelling technical systems will therefore be to failure-mode analysis for these systems, through charting the diversity of the roles filled by people in them and tracking potential conflicts between these roles. Mapping the potential failing of a technical component requires different considerations from mapping the potential 'failing' of a human component. There is a fundamental difference between a person dropping out of a role and a technical device—a pin, screw, rivet, weld—giving up, i.e., tearing, breaking, snapping, and what have you. Even a slave can choose to drop out, next to collapsing from sheer exhaustion. Anticipation of the many forms in which people can fail to comply with what is expected from them by 'the system' must rely on an entirely different conceptual and methodological apparatus compared to what we rely on for anticipating the failing of technical components. Designed devices can be formed and modified almost ad infinitum to show the one type of behaviour required for the functionality of the overall device. People cannot be similarly designed and formed; they can at most be trained for their role. If they fill a role, they do so from an individuality as a (rational) person which stays outside of the system; on top of this they may fill other, potentially competing roles in other systems or even in the same system. Roles meant to be filled by humans are defined through rules, which the person filling the role has to follow. People do not coincide with these roles, the way electrons, in a sense, coincide with their behaviour as expressed by the relevant laws of nature. Wherever a person fills a role in an instrumental system, minimally two perspectives are co-present: the perspective of the person as a human being and the perspective of the system role, which, if all is well, coincides with the perspective of the person *qua* role-player.

Finally, the perspective on modelling systems in technology adopted here may serve to point out tensions in the division of labour behind the process of structuring our complex modern technological societies, and stimulate thought on how to deal with these tensions. Operators are generally not slaves; they are citizens, and as such subject to rules that are defined outside of the system in which they fill an operator role, and which represent, again if all is well, the perspective of the civil society of which the person is a member. These rules may conflict with the instructions defining the operator role they fill. What is more, the design of many technical systems, notably the ones for which the term sociotechnical systems was introduced in Sect. 3, builds upon the existence of specific rules that govern the behaviour of people participating in these systems. However, system designers do not control the matching of system structure and societal regulation. Not only may the system, once operational, turn out to require additional regulation, since systems typically evolve and system participants learn to adapt their behaviour to the characteristics of a system and the behaviour it generates or stimulates in other participants, but the regulation in existence when the system was designed may equally be changed, requiring matching changes in the system's make-up. Further discussions of the difficulties that sociotechnical systems generate for the traditional conceptual framework still underlying most of engineering design can be found in [11, 12].

References

1. Hughes, T.P.: Networks of power: electrification in Western society, 1880–1930. Johns Hopkins University Press, Baltimore, MD (1983)
2. Wilson, B.: Systems: concepts, methodologies, and applications, 2nd edn. Wiley, New York (1990)
3. Bunge, M.: Treatise on basic philosophy, vol. 4: a world of systems. Reidel, Dordrecht (1979)
4. Hubka, V., Eder, W.E.: Theory of technical systems: a total concept theory for engineering design. Springer-Verlag, Berlin (1988)
5. Hughes, T.P.: The evolution of large technological systems. In: Bijker, W.E., Hughes, T.P., Pinch, T. (eds) The social construction of technological systems: new directions in the sociology and history of technology, pp. 51–82. MIT Press, Cambridge, MA (1987)
6. Franssen, M., Jespersen, B.: From nutcracking to assisted driving: stratified instrumental systems and the modeling of complexity. In: Engineering Systems: Achievements and Challenges. Papers Presented at the Second International Symposium on Engineering Systems, MIT, Cambridge, MA, 15–17 June 2009, http://esd.mit.edu/symp09/submitted-papers/franssen-paper.pdf
7. Franssen, M., Kroes, P.: Sociotechnical systems. In: Berg Olsen, J.K., Pedersen, S.A., Hendricks, V.F. (eds.) A companion to the philosophy of technology, pp. 223–226. Wiley-Blackwell, Chichester (2009)
8. Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., Houkes, W.: A philosophy of technology: from technical artefacts to sociotechnical systems. Morgan & Claypool, San Francisco (2011)
9. Cummins, R.: Functional analysis. J. Philos. **72**, 741–765 (1975)
10. de Weck, O.L., Roos, D., Magee, C.L.: Engineering systems: meeting human needs in a complex technological world. MIT Press, Cambridge, MA (2011)
11. Kroes, P., Franssen, M., van de Poel, I., Ottens, M.: Treating socio-technical systems as engineering systems: some conceptual problems. Syst. Res. Behav. Sci. **23**, 803–814 (2006)
12. Ottens, M., Franssen, M., Kroes, P., van de Poel, I.: Modelling infrastructures as sociotechnical systems. Int. J. Crit. Infrastruct. **2**, 133–145 (2006)

Simulation Models of Complex Social Systems A Constructivist and Expressivist Interpretation of System Dynamics

Margarita Vázquez and Manuel Liz

Abstract We take the case of System Dynamics building of computer simulation models as paradigmatic of the activity of modeling and simulation of complex social systems. We analyze the most important features of System Dynamics suggesting a constructivist and expressivist interpretation derived from the proposals of John Searle and Robert Brandom. Three problems are especially addressed: (1) The ontological problem of realism concerning the structures postulated in the models, (2) the epistemological problem of the explanatory power of the models, and (3) the methodological charge of producing merely a kind of "patchwork" through the construction of the models. We argue that a certain combination of Searle's constructivism and of Brandom's expressivism offers very promising ways of interpreting both the sort of reality that something like System Dynamics modeling is trying to capture and the sort of peculiar explanations that System Dynamics models intend to offer. Our approach has interesting consequences regarding scientific and technological disciplines devoted to the analysis of systems that, like complex social systems, are "intentionally construed" by us. Also, it is relevant regarding any development in science and technology involving the use of simulation models as a way of making 'explicit' what is only 'implicit' in certain actions or certain expert knowledge. There are also many open problems connected with our approach. We discuss some of them at the end of the paper.

This paper has been supported by Research Projects FFI2008-01205 and FFI2011-24549 (Spain).

M. Vázquez (✉) · M. Liz
Universidad de La Laguna, Santa Cruz de Tenerife, Spain
e-mail: mvazquez@ull.es

M. Liz
e-mail: manuliz@ull.es

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_31, © Springer-Verlag Berlin Heidelberg 2014

1 Introduction

Modeling and simulation is a continuously expanding field, both theoretically and practically. This is especially true with respect to computer modeling and simulation of social systems through the technique called "System Dynamics" (hereafter SD). This is one of the most interesting and widespread strategies for building computer simulation models. We will argue for the hypothesis that SD can be taken as a paradigmatic case in relation to many of the problems involved in the activity of modeling and simulation of complex social systems.

We begin by analysing some important features of SD modeling. The construction of simulation models of complex social systems, in particular computer simulation models, tries to increase our understanding and control over a system in situations in which we do not have available theories about its structure or behaviour. In order to construe the simulation models, we need to combine some sort of intuitive expert knowledge with other more formal kinds of knowledge, mainly mathematical, and both of them with some operational knowledge capable of producing a computational object that can de explored and manipulated giving rise to some sort of virtual experience. Typically, this can be found in all cases of construction of computer simulation models of complex social systems. And it is paradigmatically present in SD modeling.

In order to get a better understanding of SD modeling, we will introduce two contemporary philosophical perspectives. They come respectively from the philosophy of action and from the philosophy of logic: John Searle's constructivist proposal about the constitution of social reality, and Robert Brandom's expressivist theses on logical truth. We will argue that a combination of these perspectives offers not only a better understanding of SD modeling in particular, but a better understanding in general of the scientific and technical activities of modeling and simulation of complex social systems. It is not usual to make connections between these two authors, John Searle and Robert Brandom. However, our aim is to argue for the plausibility of a certain combination of their perspectives. Indeed, we maintain that such a combination could offer a new starting point in the philosophical foundation of the activities of modeling and simulation of complex social systems, of which SD modeling would be a paradigmatic case.

In the first place, we focus on three points of Searle's constructivist proposal: (1) the mechanisms that construct social reality, (2) the objective/subjective distinction, and (3) the implicit and unconscious nature of that construction. In a recursive way, collective intentionality, the assignment of functions and constitutive rules make it possible to construct "implicitly" all the social and institutional realities. In the second place, we focus on Brandom's expressivist theory of logic, and on his idea of a form of rationality based on the process of making conceptually "explicit" what is "implicit" in what we do. From the expressivist perspective offered by Brandom, logic could be seen as a set of expressive recourses for "saying" something about what we "do" when we make inferences. Through that process, we are placed in a position to rationally change and improve our inferential practical mastery. As we have said, we will argue that a certain combination of Searle's constructivism and of Brandom's expressivism offers very promising ways of interpreting both the sort of reality that something like SD modeling is trying to capture and the sort of peculiar explanations that SD models intend to offer.

The last point will be illustrated by discussing three crucial problems involved in the validation, or epistemological justification, of SD models: (1) the ontological problem of realism with respect to the structures postulated in SD models, (2) the epistemological problem of the explanatory power of SD simulation models, and (3) the methodological charge of merely producing a, let us say, "patchwork" when building SD models. From our discussion, we will obtain some very relevant general conclusions with respect to the interrelations among simulation models (as some sort of peculiar objects: graphics, mathematical equations, computer programs, etc.), the real systems that are intended to be modelled (a certain part of the real world), and the knowledge (both intuitive and formal knowledge) involved in the modeling processes.

In rough terms, the picture is the following one. Complex social systems are real systems intentionally construed in an implicit way out of the beliefs, desires, etc., (the mental models) of the agents involved in them. There is no other "patchwork" but the one generated by the intentional constructions of the social systems modelled. Using some mathematical and computational tools, and certain expert knowledge coming from the mental models of some subjects involved in the social systems under consideration, SD simulation models try to make explicit both the structures and dynamic consequences implicitly present in those social systems. Formally, SD simulation models are dynamical systems showing how a set of variables change over time depending on some parameters connected with them in complex ways. Usually, SD models are realised in computers. They become computational objects. More concretely, they are constituted by an equivalence class of programs offering a variety of possible behaviours in relation to a certain sort of structure. As a result of the experimentation and manipulation of those SD simulation models, the mental models of the users, enriched by those SD models, produce a very special kind of self-conscious rational improvement in the decision making processes that motivated the building of the SD models. In that way, SD models are capable of improving knowledge and rational control.

Beyond SD modeling of complex social systems, our approach has interesting consequences in two fields. First, it is relevant for other scientific and technological disciplines devoted to the analysis of any kind of systems that, like social systems, are (at least in part) "intentionally construed" by us. Second, our approach is also relevant for other developments in science and technology involving the use of simulation models as a way of "making explicit" what is only "implicit" in certain actions and expert knowledge.

2 SD as a Paradigm

System Dynamics was created by Jay Forrester, at the MIT, in the 1960s. It offers a set of tools for building computer simulation models of complex systems crucially involving feedback structures, delay structures, control structures, etc. The main field of application has been the analysis of socio-economic systems, but it has been used to address practically every sort of system in which there are strong interrelations between some variables and parameters so that one cannot study those variables and parameters independently.[1]

SD models are very good at displaying the dynamical consequences of the, in most of the cases only intuitive, structural knowledge that we have about the real systems, especially socio-economic ones, in which we are involved. This expert knowledge constitutes what is called a "mental model". The model building process in SD consists of combining that knowledge with other more formal and operational kinds of knowledge obtaining a set of computer simulation models.

The SD modeling process goes through the following interrelated stages:

1. Clarification (in ordinary language) of the mental models of some agents involved in the real systems that are intended to be modelled, usually socio-economic ones. Many times, this requires group work, the use of dialogical techniques, etc.
2. Verbal representations (also in ordinary language) of the main structural components of the system. In particular, verbal identification of the causal influences and causal loops, implicit in the mental models.
3. Iconic representations of the system putting emphasis on those components (causal diagrams, flux diagrams, the so called "Forrester diagrams", etc.).
4. Mathematical representations of the real system (using the formal tools of mathematical analysis) as some kind of dynamical system.
5. Use of computer programs (some available software) capable of generating different dynamical representations over time of the variables considered in the models. These representations (in a more or less numerical, graphical or iconic format) can be properly called "the SD simulation models" of the system.
6. Qualitative analyses of those simulation models. In particular, analyses of the sensitivity of the model with respect to changes in the values of the structural parameters.
7. Experimentation of the behaviour of the computer simulation models in relation to different courses of action over the variables and parameters of the model in different scenarios. The result of these explorations and of the manipulations of the model is a very peculiar kind of "virtual experience".

The starting point of any modeling process is a problematic situation concerning the behaviour of a certain system and a set of mental models as the only reliable source of structural knowledge. The final aim of the construction of

[1] Some important references are [1, 10, 19, 23, 26, 27] and [28].

models is to get appropriate predictions and control over the behaviour of a system. And the way of achieving that aim is by postulating a certain structure from expert knowledge. The mental models of some subjects provide that expert knowledge. Mental models come from some engaged perspectives. In the modeling process, we try to express adequately those perspectives in a verbal way, and we try to transform them into a variety of representations, both of a non-conceptual and of a conceptual kind (iconic representations, mathematical representations, computer simulation models, etc.) that can be explored in an easy way.

At this point, there are several important features in SD modeling that we want to stress. The first one is that SD modeling is especially useful when we do not have available or sound theories that can be directly applied to the systems under consideration. As we have said, it is typical that the only reliable source of knowledge is the mental models of some subjects.

The second important feature is that SD modeling rarely produces a single model. Usually, simulation models come in big families, organised in different formats and having multiple versions. With respect to any given real system, we always have a variety of different versions of verbal models, iconic models, mathematical models, computer simulation models, etc. All of them in close interaction could be considered models in relation to a certain real system.

The third remarkable feature shown in SD modeling is that models usually have a very unstable nature. They change in close connection to changes in the relevant expert knowledge, they can also change in relation to the decisions adopted concerning the boundaries of the system modelled, in relation to the application of the models to other real systems, etc.

However, and this would constitute a fourth feature, in spite of the highly unstable nature of our models, SD modeling processes offer a peculiar set, or "toolbox", of very relevant basic structures and modes of behaviour (in SD, they are called "archetypes") that can be found in many different sorts of systems in very different contexts.

The fifth feature is the following. Those basic structures, or archetypes, have both non-conceptual and conceptual components. And it is through the interaction of both kinds of components with our mental models that our ways of thinking about the real systems, and our ways of seeing them and acting upon them, change.

There is a sixth feature in close connection with the last one. The notion of mental model is involved not only in the first steps of the construction of the SD models, but in the use of those models. It is through a change in those mental models that SD models achieve their peculiar agentive power in the processes of decision making. SD models cannot be useful unless they interact with the points of view of their users, becoming an integral part of them.

As a consequence, we can point to a seventh feature. SD modeling creates a very peculiar kind of experience beyond ordinary experience and scientific experience. From the interaction with simulation models, we obtain a kind of, let us say, "virtual experience". That kind of virtual experience has both non-conceptual and conceptual contents. Our points of view about the systems modelled are enriched by those new contents, and that way our decision-making procedures are improved.

In one way or another, all the above mentioned features can be found in all areas of modeling and simulation of complex social systems. And this is what makes plausible the hypothesis that SD is a paradigmatic case in the field of modeling and simulation. The non existence of available theories is clear in relation to complex social phenomena. And the important role of mental models as a necessary source of knowledge is closely related to the need for the simulation models to interact with them, both conceptually and non-conceptually, in order to guide the decision-making processes and courses of action through the systems that are being modelled. Other features like the plurality of models with respect to any system, the lack of stability of those models, or the existence of certain sorts of basic structures are also widespread in all areas devoted to the construction of simulation models of complex social systems.

3 A Constructivist and Expressivist Approach

There are many theoretical analyses that have tried to find a philosophical foundation for SD modeling. But the rule has been to look for such a foundation through the perspectives of what can be called "general philosophies of science". Sometimes, for instance, it has been suggested that a certain kind of moderate realism, such as that of Poppers falsationism, or critical rationalism, could fit very well with SD procedures[2]. Other times, it has been maintained that some kind of practical relativism, a contextualist and pragmatist philosophy of science along the lines of Kuhn, would offer an adequate framework for the justification of the claims of SD models[3]. We ourselves, some years ago, tried to show in several papers[4] that more recent epistemological proposals, such as the "internal realism" of Hilary Putnam, are in a better position to deal with many of the conceptual problems involved in SD modelling.

We do not want to criticise the merits of those approaches here. In particular, we continue to maintain the relevance of Putnam's position as an important "third way" between realism and relativism. Indeed, a position close to Putnam's "internal realism" could deal very well with many significant features of SD. In particular, it could make sense of (1) the important role that mental models, in the sense above introduced, play in SD modeling, and (2) how some important kinds of justification, explanation and understanding are possible in spite of the plurality of alternative SD models with respect to any given system. In any case, all these sorts of general philosophical perspectives are too unspecific, and therefore their

[2] See, for instance, [5, 6] and [11].

[3] See [2, 3, 13] and [14]. For a pragmatist perspective close to Peirce, see [4].

[4] [18, 31, 32] and [30].

relevance for something like SD is very limited. Moreover, the latest developments in the field of general philosophies of science are very repetitive. They offer no significant advance. And they give a strong feeling of blockage or stagnation.

What we want to do here is to suggest a different source of insight. It is based on the philosophies of the special sciences. These philosophical perspectives are not general philosophies of science applied to a certain scientific or technical discipline. Philosophies of the special sciences try to find their own way through a great variety of resources. In particular, they elaborate their conceptual frameworks paying close attention to the fundamental notions of their respective subject matter. This is what has happened with philosophical areas such as the philosophy of mathematics, the philosophy of logic, the philosophy of biology, the philosophy of psychology, the philosophy of economics, etc. These philosophical areas do not repeat principles and norms obtained from a general philosophy of science. All philosophies of the special sciences are to a large extent autonomous and opportunistic.

As we have said, the role of a general philosophy of science is very limited in relation to our subject matter. We need a philosophical perspective capable of confronting the peculiar problems that arise in the field of modeling and simulation. In order to do this, we have to consider what are the fundamental notions in that field. And according to our hypothesis, SD could offer the clue.

In that spirit, let us introduce a couple of recent approaches in the philosophy of action and in the philosophy of logic: John Searle's constructivist proposal about the constitution of social and institutional realities, and Robert Brandom's expressivist theses concerning the justification of logical truths.

3.1 The Constructivist Proposal of John Searle

John Searle is a very well known contemporary philosopher. Until the 1980s, the majority of s work was devoted to systematising and clarifying the Speech Act Theory, one of the most important approaches in the fields of Linguistic and Pragmatics, and the most direct alternative to Chomsky formalism. Since the 1980s, Searle has become increasingly interested in aspects of the philosophy of mind. His rejection of the identification of minds with computer programs has provoked some of the most important discussions in that field in the last decades. From 1995, however, Searle has become more and more interested in certain problems of social ontology and his ideas have again been very thought provoking and stimulating.[5]

[5] [21] is the main reference about these topics. Other references include [12, 15–17, 22], and [24]. For a similar view, see [29].

We will focus on three points of Searle's constructivist proposal: (1) the mechanisms that construct social reality; (2) the objective/subjective distinction; and (3) the implicit and, many times, only unconscious nature of that construction.

According to Searle, our intentionality is an irreducible biological feature, and both social phenomena and language are manifestations of our intentionality. Social reality is a result of collective intentionality. The bearers of collective intentionality are always individual subjects, but collective intentionality necessarily involves other subjects, apart from oneself, as agents of the actions. Hence, collective intentionality can exist without the need to postulate collective subjects. In other words, we can simultaneously maintain "methodological individualism" and take into account collective intentionality as the basis of social phenomena. Social phenomena are not merely an aggregate of individual intentions and actions. Social phenomena display collective intentionality. But it is possible to give sense to the notion of collective intentionality without being engaged in the existence of collective subjects.

Collective intentionality is enough for the existence of social phenomena. However, social phenomena include much non-human behaviour. What is peculiar about human beings is that we inhabit a world full of social institutions. Searle analyses in detail two basic kinds of mechanisms used to construct social institutions: (1) The assignment of functions, and (2) the role of constitutive rules.

The kinds of functions that are important here are those that we impose on reality. They have to do with our uses or practices. We attribute to some objects functions that they are not able to perform only by virtue of their physical composition. The function of a one dollar bill, for instance, is a clear example of this. Other very important examples include the functions of representing, symbolising, etc., or, more generally, of meaning something. In all of these cases, the function requires that we accept that certain objects have a special status.

Now, let us clarify the notion of constitutive rules. Searle insists that institutional facts only exist in relation to systems of constitutive rules. Regulative rules give structure to practices that already exist, whereas constitutive rules make certain practices possible. That is, they are practices that would not exist without the rules. Chess, for instance, is constituted by the rules of chess. The formal structure of a constitutive rule is always the following:

- In context c, something x counts as y.

The function of "to count as y" would be an attributed function that x cannot perform simply by virtue of its physical composition or structure. It needs the collective intentionality of accepting that, given an appropriate context c, something x has the required status to count as y. Things like money, stamps, credit cards, rituals, conventions, etc., are what they are thanks to certain systems of constitutive rules.

Along with social phenomena, Searle considers language as an irreducible manifestation of our intentionality. Moreover, language is one of our most important social institutions. And thanks to the functions enabled by language and to the constitutive rules of language, we are able to construe all kinds of institutional phenomena.

Furthermore, a very important case in which language is capable of creating institutional facts is that of declarative speech acts: marriages, baptisms, war and peace declarations, resignations, legal sentences, contracts, certifications, etc. Here, some institutional facts linked to our linguistic practices make possible the creation of other institutional facts. We could say that our social world is full of things that "we do with words".[6]

All institutional and social reality is supported by our intentionality, which in turn is a brute, irreducible feature of our brains. Our "second" nature is a product derived from our "first" nature. Brute facts of physics, chemistry and biology are fundamental. But, beyond them, there are also social and institutional facts like marriages, wars, money, artificial objects, science, culture, economic relations or political systems. These phenomena exist because they are generated or construed by our intentionality, a brute and irreducible fact of biology. In that sense, those phenomena would be ontologically subjective. They crucially depend on our subjectivity. They are construed by us. However, there is another sense in which social and institutional phenomena are epistemologically objective They are epistemologically objective in the important sense that what we can say about them is not merely a matter of arbitrary opinion, taste or preference.

Hence, the objective/subjective distinction has two very different meanings, one of them ontological and the other one epistemological. Brute facts of physics, chemistry and biology are epistemologically objective and ontologically objective. Social and institutional facts are epistemologically objective but ontologically subjective. Sometimes, Searle also says that whereas brute facts of physics, chemistry and biology are "intrinsic features of the world", social and institutional facts are "relative to the observers". All features relative to the observers are ontologically subjective, but some of them are epistemologically objective.

In a recursive way, collective intentionality, the assignment of functions and constitutive rules make it possible to construct the social and institutional world in which we live. But, and this is also a very important point, that construction is rarely explicit and conscious. Usually, the intentional processes of construction of the social world are only implicit and unconscious. We do not need to be explicitly conscious of the ways in which social and institutional systems are construed. We only need to be equipped with certain capacities, skills, abilities and dispositions. This is what Searle calls the "Background".

[6] One of the most direct critics of Searle's social ontology is Dan Sperber. See, for instance, [25]. The main charge is that Searle atributes a causal role to what are mere Cambridge properties. According to Sperber, declarations (declarative speech acts) do not create institutional facts in any causal sense. Using Sperber's example, they would be like the Cambridge property of being "Jones' survivor" exemplified by the rest of the world when Jones dies. We could reply to Sperber's objection in an equally direct way saying that there is no reason why declarations cannot create institutional facts by means of complex causal paths.

The Background operates exactly in the same sense in which we do not need to explicitly know the constitutive rules of our natural languages. Indeed, we know how to speak the languages we are able to speak. But that knowledge constitutes our implicit mastery of those languages. With the construction of the social and institutional worlds we could say exactly the same. There is also a kind of unconscious implicit mastery. Typically, we do not construe those worlds either consciously or explicitly.

3.2 Robert Brandom's Expressivist Theses

The philosophical perspective of Robert Brandom is no less ambitious than Searle's. Brandom offers many relevant and powerful insights about profound issues concerning the philosophy of language, philosophy of mind, epistemology, metaphysics and logic. In particular, he explains the nature of meaning and the structure of the conceptual in new and fascinating ways. In addition, the same could be said of his emphasis on the importance of pragmatic norms in thought and action. As we are going to see, those topics entail certain theses about the problem of the epistemic justification or validation of logical truths that are of great interest to us.[7]

We will focus on three points of Brandom's perspective: (1) his views on meaning, (2) his expressivist theory of logic, and (3) the idea of a form of rationality based on the process of making conceptually explicit what is implicit in our practices.

Brandom's views on meaning or semantic content, both in thought and language, are pragmatist, inferentialist and contextualist. Meaning is defined in terms of the use of symbols, and the relevant uses are defined in terms of inferential practices in a public, social context. According to Brandom, representational features like reference, truth conditions, etc., applied to mental contents, i.e., the semantic contents of our thoughts, depend on the representational features of the linguistic contents present in our public languages. And these representational features of our public languages are derived from the public inferential uses of certain symbols according to the normative constraints of a certain social context. In other words, all meaning and semantical content are grounded in normative pragmatic compromises concerning the inferential use of certain linguistic items in a social context.

The relevant normative features are those that can make some asserted symbols "count as" a reason for or against other claims[8]. Hence, the semantic contents of

[7] [7] and [8] constitute the main references for this approach. He has elaborated his views in a more historical perspective, taking into account the philosophical tradition from Descartes to the present. This can be found in [9].

[8] We deliberately use the expression "count as" with the intention of suggesting some connections between Searle's and Brandom's approaches.

assertions are taken as basic, and they are defined by the inferential roles involved in our linguistic ability of giving and asking for reasons concerning those assertions. All the representational features of languages and minds are derived from that practical inferential ability. Things like logic, in a broad sense, have the expressive role of making explicit the normativity present in the inferential relations implicit in that inferential ability.

Brandom's approach is completely opposed to what has been the current representationalist paradigm. Representationalism would consider features such as reference or truth conditions applied to mental contents or to linguistic contents as primitive. According to Brandom, the representational paradigm has been ubiquitous in Western philosophy ever since the Enlightenment, and it is not easy to imagine other alternatives[9]. One alternative line of thought, however, is present in Romanticism. As opposed to the Enlightenment image of the mind as a "mirror", Romanticism proposed the image of a "lamp". Mental activity is understood not as a passive representation, but as an active revelation, full of creative and experimental ingredients. The basic picture used by Herder, for instance, is the process by which the "inner" becomes "outer" when a feeling is expressed by a gesture. In more complex cases, our attitudes are expressed in all sorts of actions, including verbal behaviours.

Brandom proposes analysing all those complex cases of expression as a matter of making explicit, in a conceptually articulated way, what is implicit in our practices. To make explicit is to turn something we initially only "do" into something we can conceptually "say". It is a process of converting a "knowing-how" into a "knowing-that", and this entails conceptualisation and re-conceptualisation. Now, once concepts are applied, we can make assertions on what is only implicit in our practices. These assertions are the sort of things that can enter as premises or consequences in our inferences and reasonings. This would open the door for a reflective understanding and a rational revision of our practices and their normative components.

Brandom's approach has powerful implications for the philosophy of logic. And we focus on that point because it has important implications for us as well. The standard way of understanding logic is as giving us access to very peculiar kinds of ideal truths: logical truths. From the expressivist perspective offered by Brandom, logic is understood in a very different way. Logic is seen as a set of expressive resources for "saying" something about what we "do" when we make inferences.

Logic would make explicit something that is implicit in our discursive inferential practices. Logical vocabulary (logical constants) serves to make that know-how explicit. The use of logical vocabulary allows us to explicitly say what we implicitly do when we apply certain concepts or when we infer some claims from other ones. Logical vocabulary allows us to make explicit the implicit inferential commitments, and entitlements, that articulate our speech acts and our

[9] About that, see [9].

thoughts. And this is the only source of epistemic justification or validation of logical truths. Logic would not be describing any ideal realm of “logical truths”.

Through the process of making conceptually explicit what is implicit in our inferential doings, we get an important kind of conceptual “self-consciousness”. Furthermore, we are then placed in a position to rationally change and improve our inferential practical mastery. Brandom calls this kind of reflective rationality “expressive rationality”.

4 Three Problems

We have presented two recent philosophical perspectives: Searle’s constructivist perspective on social and institutional realities and Brandom’s expressivist perspective on the pragmatic justification or validation of logical truths. Perhaps Searle’s perspective is not a complete philosophical account of everything that is involved in the social world, and perhaps Brandom’s falls short of being an adequate account of logical normativity, but despite these shortcomings, they offer important insights. Moreover, a certain combination of both perspectives could be very useful for us in the effort to get a reflective conceptual understanding of SD modeling.

We will try to show that by briefly analysing three crucial problems involved in the validation of SD models: (1) the ontological problem of realism concerning SD models, (2) the epistemological problem of the explanatory value of SD models, and (3) the methodological charge of merely creating a sort of “patchwork”.

4.1 The Ontological Problem of Realism Concerning SD Models

Our first problem can be introduced through a direct question: In what sense can the structures postulated by SD models be assumed to exist objectively in reality?

Here, we are faced with something that can be called an “ontological problem of realism”. Indeed, worries about the danger of a lack of realism are very frequent in the literature of SD. Moreover, sometimes the instrumental value of SD models is emphasised in a way that tries to avoid this problem by directly embracing some kind of non-realism. However, this cannot be avoided. Even though we accept that SD models have an unquestionable instrumental value, we cannot avoid this ontological problem of realism. In order to be practically effective in decision-making processes, SD models must capture those aspects that, from the points of view of the users of the SD model, are able to connect with their purposes with sufficient fidelity. So, even though SD models are built looking for a solution to a practical problem, with no other theoretical interest, we could ask the following ontological questions: In what sense can the structures postulated by SD models be assumed to exist objectively in reality? In what sense could they be said to be "real"?

Now, using the constructivist and expressivist perspectives of Searle and Brandom presented above, we can offer a highly plausible answer to this problem. In many cases, it would be completely adequate to pretend that the structures postulated by SD models represent, or describe, something objectively real in the modelled systems. SD models make explicit something that is implicit in the social and institutional worlds. These worlds are construed by us. We give them structure and reality. Therefore, the structures we make explicit in the SD models, trying to explore their dynamic consequences, can be as real as any of our intentional constructions. They can have exactly the same kind of reality as any product of our intentional actions.

The two meanings of objectivity that are distinguished by Searle are especially appropriate for dealing with our problem on a conceptual level. Many times, the structures explicitly postulated by SD models are those that implicitly give structure to the modelled systems. They are ontologically subjective. They depend on us. We construe them through our collective intentionality, by assigning a variety of functions to them, and by creating systems of constitutive rules. Even implicitly and unconsciously, we are the source of their ontological reality. However, they are also epistemologically objective. There are facts about the matter that can prove or disprove the truth of our claims over the real systems modelled. Our claims are not an arbitrary matter of taste or subjective preference. However, those structures are not ontologically objective but ontologically subjective. As we have said, they depend on us. We construe them. Even implicitly and unconsciously, we are the source of their ontological reality.

Furthermore, the subjective ontological nature of those structures offers a very simple and plausible explanation of why they cannot be easily reduced to more basic or primitive facts, described by theories which do not include the subjects involved in the social systems modelled. The reality of those social structures depends on the intentionality of the subjects involved in them. Hence, their reduction to something ontologically objective would have to entail the total ontological reduction of our subjectivity to something objective.

4.2 The Epistemological Problem of the Explanatory Value of SD Models

The second problem we want to discuss is epistemological. It has to do with the validation of our knowledge claims. How can SD models have some explanatory value? Assuming that causal explanations have a clear explanatory value, and that SD explanations are very often expressed in causal terms, that question could be formulated in the following way: In what sense are SD explanations genuine causal explanations?

Together with the worries about realism concerning SD models, there are also many discussions and analyses of the use of a causal language in SD modeling. In fact, both problems affect many other disciplines in the social sciences. In them, causal language is always suspected of being illegitimate. Very often, it is said that this causal language can only have a methodological, metaphorical or rhetorical meaning.

Again, by combining Searle's constructivism and Brandom's expressivism, we think that a plausible answer could be given to this epistemological problem. Applying mathematical and computer tools to certain expert knowledge, SD models can make explicit the structures implicit in the social system modelled. Moreover, they are especially able to make explicit the dynamic consequences of these structures. Explanations based on this explicit knowledge reveal what is only implicit in the actions of the agents involved in the modelled systems. These explanations identify some of the implicit causal relations built into the social and institutional realities that the agents have intentionally construed. From this point of view, SD explanations are no more than a further step in the process of making conceptually explicit something that is implicit in our practices.

SD explanations are similar to the explanations given when we ask what somebody did. Explaining what somebody did makes explicit relevant features of what was implicit in their doing it. Similarly, SD explanations consist in formulating relevant explicit consequences obtained from the explicit structural and dynamic knowledge offered by the SD model which, in turn, is built from what is implicit in the social systems modelled. What is implicit, and what is intended to be extracted from certain expert knowledge, are the intentional structures that the agents involved in the social systems impose on brute physical reality. These structures and their dynamic consequences are real. They are real thanks to the intentional actions of the agents. The agents are guided by a certain collective intentionality able to attribute and recognise functions, and able to accept systems of constitutive rules. That reality is epistemically objective and ontologically subjective. And SD explanations make it explicit.

Causal relations are included in the modelled social and institutional systems from the onset. Without them, these systems would not exist at all. These relations are made explicit by SD causal explanations. Because of this, the most important component of SD causal explanations is the relation between the implicit and the explicit, especially with respect to the dynamic consequences of the structures imposed by the subjects in the systems modelled. SD causal explanations try to make them as explicit as possible in order to improve rational decision-making processes.

4.3 The Methodological Charge of Merely Creating a Sort of "Patchwork"

Our last problem is methodological:. How can SD models help to elaborate deep and well founded theories about the social phenomena modeled? In particular, what would be the theoretical role of the "generic structures" (feedback structures, delay structures, control structures, etc.) used in the construction of SD models?

This is a crucial problem for the methodology of SD. Sometimes it is said that SD modeling creates a curious opportunistic "patchwork", without any theoretical orientation. This accusation would be especially relevant in relation to the integration of different kinds of SD models, and also in relation to the process in which expert knowledge is transformed into the "generic structures" analysed in SD literature.

Brandom's expressivist perspective is particularly well suited to deal with this problem. We could make a revealing and powerful analogy between Brandom's treatment of logical truths and the way that "generic structures" in SD modeling are usually employed. According to Brandom, logical relations expressed in our languages make explicit the sorts of inferences we are implicitly committed to in our inferential discursive practices. Logical relations "say" something about what we "do" when we are engaged in discursive practices. From that pragmatic basis, we can understand logical truths as those logical relations able of maintaining a peculiar constancy. Claiming that something has the status of a logical truth is to be committed to the discursive fact that, maintaining certain words constant, we can substitute the other words any way we want. The words that we can maintain constant become "logical constants". The other words become "nonlogical vocabulary".

Hence, logical truths serve to identify certain inferential constants in our discursive practices. At this point, we could apply Brandom's approach to the methodological problem we are discussing. Exactly in the same way in which we can say that logical truths serve to identify certain constants in our inferential discursive practices, we could say that "generic structures" in SD serve to identify other sorts of constants in our social practices. We would obtain a kind of expressivist conception of the "generic structures" of SD, quite analogous with Brandom's expressivist conception of logical truths.

It is important to note that from the perspective we are proposing, the lack of a previously defined theoretical orientation is not a defect but, in a certain sense, a virtue of SD modeling. In the same sense that the lack of a previously defined theoretical orientation is not a defect but a virtue in the discovery of logical truths. In this case, a previously adopted theoretical orientation could introduce important mistakes in the evaluation of our inferential practices. In a similar sense, a previously adopted theoretical orientation in the building of SD models of social systems could introduce important mistakes in the evaluation of the social practices from which social systems are built.

Something very close can be said of the use of the same SD "generic structures" in different contexts, for instance when we are modeling very different social phenomena. Those structures would be "generic", very basic, ones simply because they can be used that way, with a different particular content in each case. They reflect or express something that is implicitly present in many of our actions. Exactly in the same sense in which logical truths would reflect or express something that is always implicitly present in our discursive practices.

Of course, the normative force of SD "generic structures" is not so strong as the normative force of something like logic. Logic has a maximum of normative force. Anyway, the important point would be that in both cases we have some necessities and possibilities far beyond the necessities and possibilities found in the natural worlds of physics, chemistry and biology. Additionally, in both cases the source of such "second nature" normativity would be the intentionality displayed in our actions. Moreover, in both cases we get a very special kind of conceptual "self-consciousness" by means of which we can rationally improve our prac.

4.4 Interactions

So far, we have identified and discussed three important problems and we have tried to show that a certain combination of the philosophical perspectives opened by Searle's constructivism and Brandom's expressivism would offer very powerful conceptual recourses to deal with them.

The basic idea is that in SD modeling of complex social systems we explicitly obtain what we implicitly put into the systems modelled. The social and institutional systems modelled are implicitly construed by us. And SD modeling makes their structure and dynamic consequences explicit in order to achieve a better "self-conscious" rational position in decision-making processes.

Furthermore, from the new constructivist and expressivist approach suggested, the interrelations among mental models, complex social systems, and SD models (a classical topic in SD reflections) can be understood in a very simple and clarifying way:

1. Complex social systems are real systems intentionally construed in an implicit way out of the mental models of the agents involved in them. They are real systems that are epistemologically objective but ontologically subjective.
2. Using certain mathematical and computational tools, and applying them to certain expert knowledges, SD models try to make explicit the structures and dynamic consequences implicitly present in those complex social systems.
3. Mental models enriched by SD models, with all of their non-conceptual iconic components, get a special kind of "self-conscious" conceptual qualification. And this makes possible a rational improvement of the relevant decision-making processes that inspire the building of SD models.

Complex social systems are real systems about which objective knowledge is possible. However, their reality has a subjective ontological source. We construe them. SD models make explicit some of those constructive components and their dynamic consequences. In this process of making conceptually explicit something that is implicit in our actions, SD modeling needs the help of some expert knowledges. Mathematical and computational tools are also crucial because they constitute the expressive recourses capable of making conceptually explicit the relevant constructive components and dynamic consequences. Without those recourses, we could not make them explicit. And they do it in ways that make possible the existence of completely new empirical fields (some kinds of "virtual experiences").

The last point is very important. And it is connected with the fact that what is made conceptually explicit in SD models must intimately interact with the mental models of the users of the SD models in order to improve their decision-making processes. These decisions are part of the constructive components and dynamical consequences of the complex social systems in which the subjects are involved. They are part of their practice as agents. And SD models cannot be useful unless they are finally integrated with the implicit forces that construe the social systems modelled.

5 Conclusions and Open Questions

Let us summarise our main results. We began adopting the hypothesis that SD modeling offers a paradigmatic case of the activity of modelling and simulation of complex social systems. Trying to find some sort of reflective understanding of SD modelling, we suggested that instead of merely applying some sort of general philosophy of science we should employ perspectives that are more sensitive to the peculiarities of SD modelling.

The analysis of some of the features present in the social systems modelled and in the SD modelling of those systems leads us to the fields of the philosophy of action and of the philosophy of logic. On the one hand, social and institutional systems entail the implicit construction of objective realities through our intentions and actions. On the other hand, SD modelling of those systems is built with the aim of rationally improve explicit decision-making processes. In order to continue this analysis, we introduced two recent philosophical approaches: John Searle's constructivist perspective on the constitution of social reality, and Robert Brandom's expressivist theses on the justification of logical truths. According to Searle, social and institutional phenomena are construed through the recursive iteration of three basic mechanisms: collective intentionality, the assignment of functions, and systems of constitutive rules. The social and institutional phenomena are epistemologically objective, but ontologically subjective. Moreover, in general we construe them only in an implicit and unconscious way. According to Brandom, logic does not describe or represent any ideal realm. It has an expressive role linked to what is implicit in our inferential discursive practices. Logical vocabulary serves to make explicit what is implicit in our inferential commitments, and logical truths express some invariances present in them. The justification of logic is pragmatic. However, it is not based on mere success. Through the process of making explicit what is implicit in our actions, we get a very important kind of conceptual "self-consciousness" able to rationally improve our inferential practices. We have argued that a certain combination of the perspectives of Searle and Brandom could be very useful to achieve a reflective understanding of SD modelling. We applied their constructivist and expressivist views to the discussion of three crucial problems: the ontological problem of realism concerning SD models, the epistemological problem of the explanatory value of SD models, and the methodological charge of merely producing a kind of "patchwork". Finally, we generalised that constructivist and expressivist approach analysing the interrelations between mental models, complex social systems, and SD models.

Beyond SD modelling of social systems, our approach has interesting consequences for two fields. Firstly, it is relevant to other scientific and technological disciplines devoted to the analysis of any sort of system that, like social and institutional systems, are at least in part intentionally construed by us. Secondly, assuming that SD modelling is a paradigmatic case of the scientific and technological use of computer simulation models in order to increase our knowledge and control, our approach would be also relevant in relation to other developments in science and technology involving the use of simulation models as a way of making explicit what is only implicit in certain actions and in certain expert knowledge[10].

There are also many open questions. We will comment very briefly three particularly important ones. The first one concerns constructivism. The construction of social phenomena has limits. It can be constrained by a variety of factors. Mainly, it can be constrained in three ways: (1) by psychological or subjective facts entailing limits to our constructive powers, (2) by objective facts in the reality outside the subjects, generally having to do with complexity, and (3) by the interrelations among those subjective and objective facts[11]. With all of this in mind, perhaps it would be better to speak of a "bounded constructivism", in analogy with the familiar idea of a bounded rationality. It would be very interesting to analyse that bounded constructivism in precise terms.

The second question has to do both with constructivism and with expressivism. Imagination has a role both in the construction of social phenomena and in the expressive move of saying explicitly what is implicit in our actions. Moreover, sometimes that role is crucial. It is so, for instance, when we have to cope with some kind of constraint, or "bounding", in the construction of social phenomena. And it is also crucial in every decision-making process guided by the explicit

[10] For some authors, computer simulation would entail a whole new way of doing and understanding science and technology. See [1] and [33].

[11] About the crucial implications of that point for SD, see [27]: Chap.1. His analysis focuses on the fact that, very often, complexity hinders evidence and our ability to discover the delayed and distant impacts of our interventions, generating unintended side effects.

structural and dynamic knowledge obtained from the SD models. Again, it would be very interesting to analyse the role of imagination in these cases.

The third open question has to do with the application of the proposed combination of constructivism and expressivism to other important problems that arise in the field of SD, as well as in many other areas of modeling and simulation. There is a strong tendency to view SD as a methodology very apt to improve learning processes in complex social systems[12]. When SD is considered from that perspective none of the usual philosophical approaches offers much help. What is needed seems to be an approach sensitive to the fact that social and institutional systems are constructed by us, and also sensitive to a certain, let us say, "Socratic" conception of learning as a move from what is implicit in our actions to what we are able to make conceptually explicit. Both things are at the core of our proposal.

References

1. Axelrod, R.: Advancing the art of simulation in the social science. In: Conte, R., Hegselmann, R., Tema, P. (eds.) Simulating Social Phenomena. Springer-Verlag, Berlin (1997)
2. Barlas, Y.: Formal aspects of model validity and validation in system dynamics. Syst. Dyn. Rev. **12**(3), 183–210 (1996)
3. Barlas, Y., Carpenter, S.: Philosophical roots of model validation: Two paradigms. Syst. Dyn. Rev. **6**(2), 148–166 (1990)
4. Barton, J.: Pragmatism, System Thinking and System Dynamics. System Dynamics Conference (1999)
5. Bell, J.A., Bell, J.B.: System dynamics and the scientific method. In: Randers, J. (ed.) Elements of the System Dynamics Method, Chapter 3, pp. 3–22, MIT Press, Cambridge (1980)
6. Bell, J.A., Senge, P.: Methods for enhancing the refutability in system dynamics modeling. TIMS Stud. Manag. Sci. **14**, 61–73 (1980)
7. Brandom, R: Making it Explicit. Reasoning, Representing, and Discursive Commitment. Harvard University Press, Cambridge (1994)
8. Brandom, R: Articulating Reasons. An Introduction to Inferentialism. Harvard University Press, Cambridge (2000)
9. Brandom, R: Tales of the Mighty Dead. Historical Essays in the Metaphysics of Intentionality. Harvard University Press, Cambridge (2002)
10. Forrester, J.: Industrial Dynamics. MIT Press, Cambridge (1961)
11. Forrester, J., Senge, P.: Test for building confidence in system dynamics models. In: Legasto, A., Forrester, J., Lyneis, J. (eds.) System Dynamics. North-Holland, Amsterdam (1980)
12. Grewendorf, G., Meggle, G. (eds.): Speech Acts, Mind, and Social Reality. Kluwer, Dordrecht (2002)
13. Homer, J.: Why we iterate: Scientific modeling in theory and practice. Syst. Dyn. Rev. **12**(1), 1–19 (1996)
14. Homer, J.: Structure, data and compelling conclusions: Notes from the field. Syst. Dyn. Rev. **13**(4), 293–309 (1997)
15. Koepsell, D. (ed.): John Searle. Special issue of, the American Journal of Economics and Sociology 62 (2003)

[12] See, for instance, [23] and [27].

16. Koepsell, D., Moss, L.: John Searle's Ideas about Social Reality. Blackwell, Londres
17. Lepore, E., van Gulick, R. (eds.): John Searle and His Critics. Blackwell, Oxford (1991)
18. Liz, M., Aracil, J., Vázquez, M.: Knowledge, Control, and Reality: The Need of a Pluralistic View in Control System Design. Proceedings of the 13th IFAC World Congress (1996)
19. Richarson, G.: Problems for the Future of System Dynamics. Syst. Dyn. Rev. **12**(2), 141–157 (1996)
20. Schmitt, F. (ed.): Socializing Metaphysics. Oxford, Rowman &Littlefield (2003)
21. Searle, J.: The Construction of Social Reality. The Free, New York (1995)
22. Searle, J.: Mind, Basic Books, Language and Society. Philosophy in the Real World, New York (1999)
23. Senge, P.: The Fifth Discipline. The Art & Practice of the Learning Organizative, Doubleday Currency (1990)
24. Smith, B. (ed.): John Searle. Cambridge Unversity Press, Cambridge (2003)
25. Sperber, D.: The deconstruction of social unreality. Seventh European Congress of the Analytical Philosophy Association, Milan (2011)
26. Sterman, J., et al.: A skeptics guide to computer models. In: Barney, G. (ed.) Managing a Nation: The Microcomputer Software Catalog, pp. 209–229. Westview Press, Boulder (1991)
27. Sterman, J.: Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston (2000)
28. Sterman, J.: All Models are Wrong: Reflections on Becoming a System Scientist. System Dynamics Review **18**(4), 501–31 (2002)
29. Tuomela, R.: The Importance of Us: A Philosophical Study of Basic Social Notions. Stanford Univ. Press, Stanford Series in Philosophy, Stanford (1995)
30. Vázquez M., Liz, M.: Models as points of view. The case of system dynamics. Found. Sci. **16**(4), 383–391 (2011)
31. Vázquez, M., Liz, M., Aracil, J.: An Epistemological Framework for System Dynamics Modelling. Revue Internationale de Systmique **9**(5), 461–89 (1995)
32. Vázquez, M., Liz, M., Aracil, J.: Knowledge and reality: Some conceptual issues in system dynamics modelling. Syst. Dyn. Rev. **12**(1), 21–37 (1996)
33. Winsberg, E.: Simulated experiments: methodology for a virtual world. Philosophy of Science **70**, 105–125 (2003)

Architectural Design Thinking as a Form of Model-Based Reasoning

Pieter Pauwels and Rens Bod

Abstract Model-based reasoning can be considered central in very diverse domains of practice. Recently considered domains of practice are political discourse, social intercourse, language learning, archaeology, collaboration and conversation, and so forth. In this paper, we explore features of model-based reasoning in architectural design and construction. Additionally, an indication is given of some existing suggestions of how model-based reasoning systems may be simulated in an automated environment. We extend these lines of thought into our own simulated environment and give indications of how such a model-based reasoning system can not only give us better insights in the architectural design and construction practice, but also why it is so hard for such a system to eventually surpass human capabilities in this area of practice.

1 Introduction

Model-based reasoning strategies are typically considered central to scientific research practice and scientific discovery. In following a model-based reasoning strategy, reasoning agents continuously switch back and forth between their own mental models of particular parts of the world and the parts of the world that surround each of them. The reasoning agents use their models of a part of the world

P. Pauwels (✉) · R. Bod
Institute for Logic Language and Computation, University of Amsterdam,
P.O. Box 94242 1090 GE Amsterdam, The Netherlands
e-mail: P.Pauwels@uva.nl
URL: http://www.uva.nl/

R. Bod
e-mail: Rens.Bod@uva.nl

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_32, © Springer-Verlag Berlin Heidelberg 2014

to make sense of the situation in front of them. At the same time, they try to verify or justify their model and strengthen it by additional empirical confirmations.

An increasing number of research initiatives acknowledges the role that model-based reasoning plays in application domains other than scientific discovery. For instance, the role of model-based reasoning, and in particular the abductive parts of model-based reasoning, has been indicated in making medical diagnoses [1], in second language learning [2], in plain collaboration and conversation [3, 4], in archaeology [5], in a political and social context [6], and so forth. We will look at the role of model-based reasoning in the context of architectural design and construction in this paper. In this discussion, we will additionally take into account the concept of motivation-biased design, as it is outlined by [7] and [8], and the concept of linkographs, as it was originally outlined in [9–11], because of the close relevance of these concepts both to architectural design and construction, and to model-based reasoning.

Additionally, we will look at some existing computational models that simulate processes similar to model-based reasoning. More precisely, we will consider the Data-Oriented Parsing (DOP) framework [12, 13], the non-omniscient agent model by [14], and our own conceptual implementation of Peirce's process of inquiry [15]. In documenting the latter, we will have a closer look at the work of Peirce that is closely related to his process of inquiry. In this overall consideration, we will particularly consider the way in which these computational models could be of relevance in a context of architectural design and construction, thereby relying on earlier work that is documented in [16].

2 Model-Based Reasoning in Architectural Design and Construction

2.1 Reasoning in Design Practice

Model-based reasoning is not only about the ways in which models are being *used* in practice and their effect on this practice, it is also very much about the way in which models are being *constructed* and continuously *revised* in the practitioner's mind. The world around the practitioner has a tremendous impact on the practitioner, in the sense that, in the long run, it forms the models that are being used by that practitioner to act upon the world. As a result, the models that are being used at a specific moment in time, for instance by a musician, clearly depend first and foremost on the world in which this practitioner has grown up and learned to appreciate certain musical styles. The same holds for scientific models, mathematical models, language models, and so forth. The notion of 'practitioners' and 'practice' is thus central to the complete idea of model-based reasoning, as it shapes the models that are being used for reasoning.

Recent theories of design thinking similarly indicate the central position of practice in design, which includes architectural design and construction, although this has not always been the case. In particular at the time when computers and automation processes were just starting to emerge, around the 1960s, many design theorists claimed that design is all about finding a solution to a well-defined problem using rational methods (see, for instance, [17, 18]). Eventually, however, many design theorists shifted their claims (see, for instance, [19, 20]) and started to see design thinking as a different kind of problem-solving, in which the problem is not well-structured but 'wicked' [21] or 'ill-defined' [22]. One of the most striking, and perhaps one of the most appropriate ways for describing the nature of wicked problems, in contrast to well-structured or tame problems, is to state that tame problems can be solved, whereas wicked problems can only be 're-solved'. *"[...] Social problems are never solved. At best they are only re-solved - over and over again"* [23] (p. 160). As a result, a model of planning (and thus of design) was more often understood *"as an argumentative process in the course of which an image of the problem and of the solution emerges gradually among the participants, as a product of incessant judgment, subjected to critical argument"* [21] (p. 138). In this model, design is not understood as a rational and methodological science, but more as a discipline, a practice or a profession. It is this model of design as a 'practice' that quite closely resembles the idea of model-based reasoning and that will be considered in this paper.

Note that other works have considered the similarity between architectural design and construction as a practice and other model-based practices. For instance, in [24] (pp. 54–55), architectural designers, baseball pitchers and musicians are considered reflective practitioners that entirely rely on a similar kind of *reasoning in practice*. This kind of reasoning is referred to as 'reflection-in-action' by [24], but we argue that one could just as well use the term 'model-based reasoning' here. The notion of *reflection-in-action* can be understood along the same lines as model-based reasoning, in the sense that practitioners continuously try to make sense of their environment and their actions in this environment using their background knowledge, and act upon their environment accordingly, eventually resulting in a refined or at least an altered understanding or background knowledge. Also the way in which design thinking is explained in the work by Lawson [25], Cross [26] and Goldschmidt [27, 28] resembles the general idea of model-based reasoning. Goldschmidt's term 'seeing as' [27, 28], for instance, is one appropriate example of such a parallel. She indicates how designers that are 'seeing as', are essentially looking at a certain situation in front of them, making sense of this situation using their own model or background knowledge (i.e. seeing as something they experienced before), and accordingly act upon the result of the sense-making process.

2.2 The Leap of Interpretation in Architectural Design and Construction

Both in model-based reasoning and in regular theories of design thinking, a significant portion of interest and focus is reserved for the initial phase of the reasoning process, namely the phase in which one comes across a new observation or empirical experience, and tries to explain or make sense of this experience. In a model-based reasoning context, abductive reasoning is typically considered central to this phase. In a design thinking context, Cross [29] refers to a useful interpretation by [30]: *"there is a prior and pervasive kind of reasoning that scans a scene and sizes it up, packing into one instant's survey a process of matching, classifying and comparing. [...] Metaphoric appreciation, as all the words we have used suggest, is a work of approximate measurement, scaling and comparison between like and unlike elements in a pattern"* [30] (p. viii). Later on, Cross [31] refers to several other research initiatives that distinguish a very similar kind of reasoning as fundamental for design thinking, thereby mentioning the terms abductive reasoning, productive reasoning and appositional reasoning as they were called by their respective inventors Peirce, March and Bogen [32–34]. Also Simon [35] refers to this prior step in architectural design and construction: *"The problem-solving process is not a process of 'deducing' one set of imperatives (the performance programme) from another set (the goals). Instead, it is a process of selective trial and error, using heuristic rules derived from previous experience, that is sometimes successful in discovering means that are more or less efficacious in attaining some end. [..] the process of derivation is not a deductive process, it is one of discovery. If we want a name for it, we can appropriately use the name coined by Peirce and revived recently by Norwood Hanson [1958]: it is a retroductive process"* [35] (p. 151).

Especially the mention of Peirce's term 'abductive reasoning'—which can be considered identical to his earlier term 'retroductive reasoning'—is interesting, not only because abductive reasoning is one of the key concepts in model-based reasoning, but also because it places creativity in architectural design and construction next to one of the key contributions by Peirce to philosophical discourse, namely the process of (scientific) inquiry. Following this latter line of thought, the design process might be understood as a kind of process of inquiry, similar to other processes of inquiry, with at its starting point a *leap of interpretation* that is constituted by an abductive reasoning step.

Understanding architectural design and construction as a process of inquiry is not straight-forward and evident, however. Peirce's process of inquiry, namely, is first and foremost a process of *scientific* inquiry: it is most often explained in the context of *scientific discovery*. Architectural design and construction, on the other hand, is typically considered to be either engineering and/or art. As a result, the 'products' of architectural design and construction are closer to *inventions* rather than discoveries. A useful discussion on the differences and similarities between inventions and discoveries can be found in [36–38]. Especially regarding the element of abductive reasoning, important remarks are made in [38]. It is argued that *"there is no reason to believe that the cognitive processes underlying questioning in the two contexts [invention and discovery] are fundamentally different. [...] roughly the same cognitive theory should be able to handle both scientific and technological questions"* [38]. At the same time, it is argued that *"by definition, abduction concerns the generation of explanatory hypotheses, so abduction is not central to invention, which primarily involves the generation of answers to questions that are practical rather than explanatory"* [38]. In short, scientific discovery comes forth from puzzling observations that pose the observer before a why-question ('*why* did X happen as it did?). These why-questions can be answered by making hypotheses, which rely on abductive reasoning. Invention, in contrast, do not come forth from puzzling observations, but rather from a need ('*how* can we make X so that it accommodates Y?'). As a result, the role of abductive reasoning in the latter is unclear and, in any case, the cognitive process underlying invention and discovery cannot be considered identical. Apart from that difference, however, there are major similarities between inventions and discoveries, as is also argued in [38]. *"Although abductive inference to explanatory hypotheses is much more central to scientific discovery than to technological innovation, inference to possible solutions to technological processes seems to involve very similar representations and processes"* [38].

For this paper, we will start from the more general perspective of the cognitive processes that are involved in architectural design and construction. The domain of architectural design and construction is hereby considered as a type of engineering that produces innovative products or inventions. The reasoning processes involved in producing design products will be considered similar to the reasoning process involved in scientific discovery and will thus include abductive reasoning. A more detailed discussion on the nature of the creative process involving interpretation within architectural design and construction and its relation to abductive reasoning would be a valuable topic for further research.

2.3 Linkographs: Quantitative Analyses of the Design Process

We will not look into concrete examples of design processes in this paper, because this requires a different focus. Diverse example studies of design processes are nonetheless available elsewhere (e.g. [39–41]). Examples exist that directly aim at indicating how abductive reasoning is in play in design thinking [42]. Most investigations of design thinking, however, make less direct cognitive analyses using protocol studies [40, 43]. In this method, a track record is obtained from designers involved in design activity through think-aloud protocols [44]. An example protocol study can be found for packaging system designers [39] (see also other example studies [40, 41]).

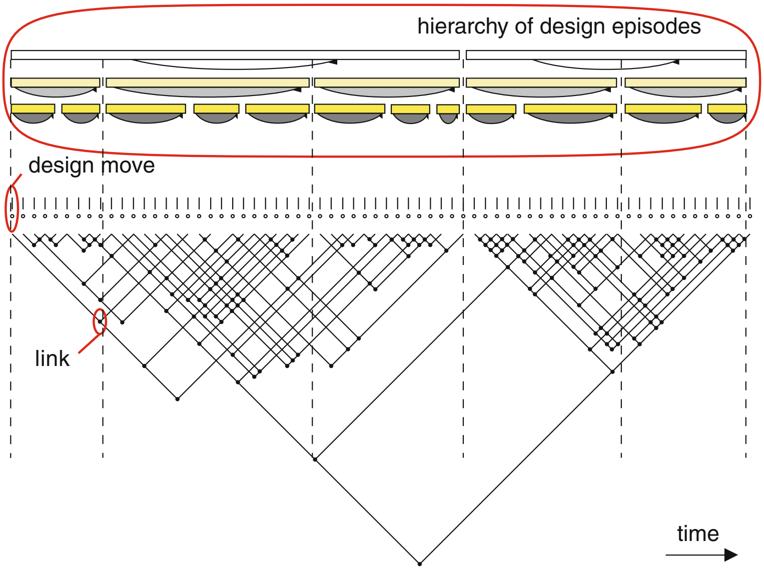


Fig. 1 An example linkograph that might be associated to a design process (e.g. making a sketch, discussing with colleague designers, thinking aloud). Note that this linkograph does not represent a real design process. It is only meant for outlining the idea behind the linkograph method within this paper

Although diverse methods exist to analyse protocol studies, linkography can be considered as one of the most successful [9–11]. An example linkograph is shown in Fig. 1. In producing a linkograph, the protocol study is subdivided in design moves, each of which is recorded in a sequential/chronological representation (see horizontal sequence of design moves in Fig. 1). Second, the protocol study is analysed for associations between the distinct design moves. These associations are represented by links in the linkograph (see diagonals in Fig. 1). The Link Index (LI) and Critical Moves (CM) parameters are then two of the most important numerical parameters that can be used to interpret the resulting linkograph [10, 11]. As such, the linkography method provides a way to make a quantitative analysis of the protocol study instead of a qualitative and thus more subjective analysis [11]. Note, however, that (subjective) expert interpretation is still needed in (1) choosing and encoding design moves, (2) choosing and encoding links, and (3) evaluating the resulting linkograph [45]. Nonetheless, the linkography method is one of the best descriptive approaches to quantitatively analyse the design process.

The purpose of a linkograph representation is of course not that it gives an exact representation of the actual reasoning process of the designer. The purpose is that it provides the means to enable representations of how the reasoning process of the designer might have occurred, or, preferably, how it most probably occurred, and to make a quantitative analysis of this process. The linkograph in Fig. 1 does not represent any specific design process, but is instead constructed specifically for the purpose of illustration and discussion within this paper only.

The crucial notion for this section of this paper is that episodes can be distinguished in linkographs and thus also in the design process (see top in Fig. 1). The notion of episodes within linkographs allows to distinguish between 'within-episode links' and 'cross-episode links', which are links in the linkograph that do or do not cross the boundaries of a specific episode. Episodes then appear to reflect distinct periods of activity in which the designer simply proceeds in elaborating a specific idea with only minor creative shifts (i.e. within-episode links). In the context of model-based reasoning, these periods can be considered as periods in which the designer maintains a specific model and continuously makes actions according to this model. The interesting points then become the periods where one episode stops and another one starts. This would namely imply a shift of model in a model-based reasoning perspective. Following our earlier arguments, these moments might thus be considered as the points where abductive leaps of interpretation are made, resulting in new ideas, new models, new understandings.

2.4 The Place of Analogy

Of course, a leap of interpretation is never randomly made. Also in the linkograph representation, links are available that span two or more episodes (cross-episode links). These links indicate that episodes remain associated to previous episodes. The question rises then how one changes a working model into a different, yet related working model. Following the previous section, this supposedly occurs through a step of abductive reasoning. Yet, there might also be a place in there for *analogical reasoning*. Similar references to analogical reasoning are also made in [36, 38] for scientific discovery and invention. Analogical reasoning allows one to find a structural alignment or mapping between a base and a target pattern residing in (partially) different domains [46–50]. By making such an analogical mapping, familiar knowledge about the base pattern can be related to the target pattern, thereby filling in the gaps of the target pattern and creating new knowledge. In a context of architectural design, analogical reasoning is often explained as occurring between a new design-related experience (building, sketch, 3D model, conversation, and so forth) and a previous design-related experience [51]. But also in the very act of sketching, analogical reasoning is crucial, because it allows reinterpreting or 'seeing as' [27, 28]. In 'seeing as', the designer reinterprets the sketch and, through analogy, adds new and original meaning to it, thereby generating new ideas [27, 28].

Suppose that the linkograph in Fig. 1, for instance, represents a sketching process. In this case, the design process has started long before the starting point of the linkograph and includes many episodes which include activities other than sketching. The cross-episode links between these episodes can then be considered as representations of structural alignments, or analogies, across episodes. In other

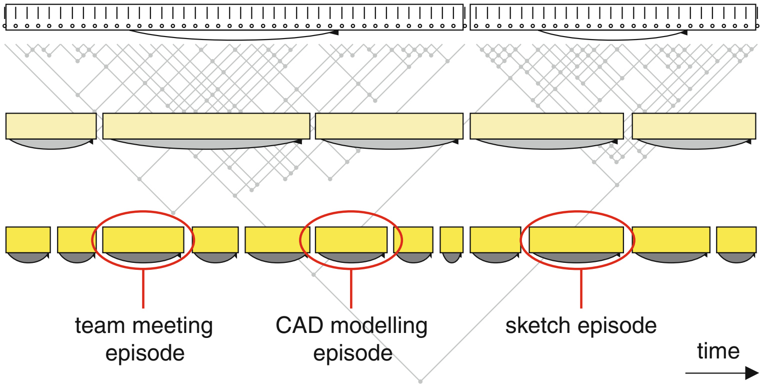


Fig. 2 An extrapolation of the linkograph principle indicates how the acts in a design process can be considered as acts within a greater sequence of design acts by one person, thereby providing a medium for representing both short-term and long-term analogies made by the designer

words, in the form of linkographs, one might have a medium that allows to represent analogies between episodes. Considering that such cross-episode links might go back for quite some time, spanning multiple design episodes, such linkographs might form a very powerful medium to investigate analogies and model-based reasoning in architectural design processes across very diverse kinds of design episodes (Fig. 2—team meeting episode, CAD modelling episode, sketch episode).

2.5 The Central Role of the Surrounding Environment

Following the idea that the environment in which the designer acts is central to the architectural design and construction process, and the twists and turns that this process takes, we outline in a bit more detail what constitutes this surrounding environment, and how this relates to the concepts of abductive, model-based and analogical reasoning that were outlined in the previous sections. As already indicated with the linkograph in Fig. 1, the architectural design process can be considered as a sequence of interrelated episodes. Of course, sketch episodes are among the most obvious examples of episodes that designers go through at design time. As Goldschmidt [27] indicates, sketches are not only visual expressions of what one wants to express; they also are elements for reinterpretation and thus for generating all kinds of new knowledge. Cross similarly refers to the importance of sketching because it enables a designer to explore several solutions and problems to a certain design situation at once, thereby considering several levels of detail at once [52]. Schön, in turn, refers to the habit of many a designer to continuously make representations of the design situation at hand in documents, plans, and sketches, thereby allowing a designer both to answer a previously generated problem situation or design situation, and to frame the design situation anew into an alternative perspective [24].

Design episodes might be entirely about sketching, but architectural design includes many more episodes which contribute to the design process, including dialogues and conversations, thought experiments, physical modelling, CAD modelling, simulations, and so forth (Fig. 2). A valuable theory in this regard is formed by Lawson's theory of so-called 'guiding principles' or 'design philosophies' [25] (pp. 159–180). Both terms can be understood as the background knowledge or the knowledge by experience of a designer, which can be considered parallel to the 'model' in model-based reasoning. These guiding principles are constituted by the episodes that a designer has experienced, and they steer this designer in one way or another, both consciously and unconsciously, depending on the current context and the content of these guiding principles. A design episode in which intense conversation occurs between designers and the design team they are part of, can consequently be considered as an episode similar to sketching (but with a different context of course), in which the designers try to make sense of the design situation and try to propose actions for altering this design situation. Similarly, also information systems (CAD modelling environments, simulation or calculation environments, archive applications, and visualisation applications) can be considered different environments in which a certain episode in the design process can take place, thereby altering both the further design process and the guiding principles of the designer (see also [16]).

According to Lawson, these guiding principles include not just objective, factual information but also information involving, for instance, motivations, beliefs, values, and attitudes. Guiding principles may be very intense and clearly structured, or they may be vague and unclear, but they always influence design decisions one way or another. In some research initiatives, they are almost considered part of a *"personal religion"*, thereby implicitly redefining design as *"a very complicated act of faith"* [53]. With sometimes profound intensity, designers hold to these personal guiding principles, believing it *"morally right"* to follow them. As such, the notion of guiding principles relates to the notion of motivation-biased design, which has been proposed in the context of model-based reasoning as well [7, 8]. In short, *"motivation-biased design concerns how positive attitudes of designers can inhibit critical evaluation of their designs"* [7]. In other words, because of the preference that designers have built over the years for certain methods, design elements, or construction companies, for instance, designers that are particularly prone to motivation-biased design tend to deploy such methods, elements or construction companies, whether they be appropriate for the particular design situation or not. If we follow our earlier indications about analogical reasoning in the linkograph example (Figs. 1, 2), the only way in which such motivations can find their way into the design process, is through analogies that are made at design time between a current (design) episode and a previous (design) episode. If these design episodes have a highly 'motivational' character, motivation-biased design can indeed occur. The cross-episode links in the linkographs, which span diverse design episodes, represent structural alignments across episodes, and can thus also represent forms of motivation-biased design over spans of time.

3 Simulating Model-Based Reasoning?

Diverse simulations have already been proposed for model-based reasoning processes, including conceptual, logical and implemented simulations. None of those simulations comes close to a simulation of a process that is similar to the architectural design process outlined above. Yet, considering that reasoning in architectural design and construction might be considered a form of model-based reasoning, such a simulation might be targeted as well. We will therefore look here into some of the existing simulations of reasoning processes similar to model-based reasoning. More specifically, we will look into the simulation of non-omniscient agents as it is suggested in logical terms by Velazquez-Quesada [14] and one of the Data-Oriented Parsing (DOP) frameworks, namely Unsupervised DOP (U-DOP), which is suggested by Bod [13]. We will not go into detail for both simulations, but we will outline how the considered simulations can be related to the previous section on reasoning in architectural design and construction, and thus indicate why such simulations can be useful in such a context and in a context of model-based reasoning in general.

3.1 Non-omniscient Agents in a Dynamic Epistemic Logic Framework

Epistemic Logic (EL) is used in many contexts and for many purposes. It is relevant in the discourse of this paper in the sense that it enables handling knowledge and beliefs in a reasoning system [54, 55]. As such, EL can enable to make a logical representation of the way in which knowledge and beliefs change throughout the episodes of a design process. For instance, the belief that a specific construction company fits better to a specific construction project than any other construction company can be expressed in EL, after which a logical system might confirm why a designer chooses to continue with a specific construction company in a specific project and context (cfr. motivation-biased design). As such, an EL system might, for instance, enable to simulate, or at least represent, the evolution of beliefs and knowledge throughout each of the (design) episodes as they were previously outlined in the linkograph (Fig. 3).

Of particular relevance for model-based reasoning are the non-omniscient agents that are proposed in [14] as part of a Dynamic Epistemic Logic (DEL) framework. In this system, reasoning agents are not considered to be omniscient, as is traditionally the case in DEL systems, but non-omniscient. This allows to

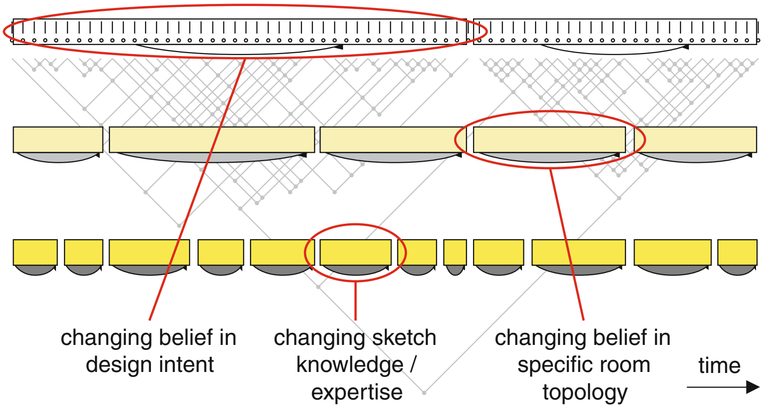

changing belief in design intent
changing sketch knowledge / expertise
changing belief in specific room topology
time

Fig. 3 Changing beliefs at different levels of the linkograph with design episodes

represent more 'real' agents, in the sense that the knowledge and beliefs used by the agent are not closed under logical consequence. Additionally, this setting puts more focus on the representation of steps for knowledge *change* and belief *change*, which would make no sense in an omniscient setting.

As is also outlined in [14], the non-omniscient agent system allows to represent 'small steps in dynamics of information'. As such, it could also represent the 'small steps of information' that are occurring at different levels in the episode hierarchy of the design process (see top of Fig. 1). More precisely, a non-omniscient agent in the DEL framework could simulate one of the levels of design episodes in Fig. 1 or 3, with the within-episode links in this episode being the small steps in the dynamics of information. Assuming that episodes can be grouped in larger episodes, some of the cross-episode links might actually become within-episode links. In this sense, the small steps in dynamics of information thus occur both on a lower level and on a higher level. As such, this mechanism presented in [14] might be very useful in getting to a simulation of the reasoning process in architectural design and construction that is documented in the first section of this paper.

3.2 Unsupervised Data Oriented Parsing

The DOP approach is primarily developed for parsing textual data and aims first and foremost at supervised textual parsing, implying that it starts from large text corpora that were human-annotated beforehand and thus provide a large set of example derivations. The DOP approach hereby typically considers a corpus of sentences stemming from realistic contexts and including, depending on the case study, baby utterances, animal sounds, adult human language, and so forth. Human annotators subsequently annotate these sentences, so that a tree structure is

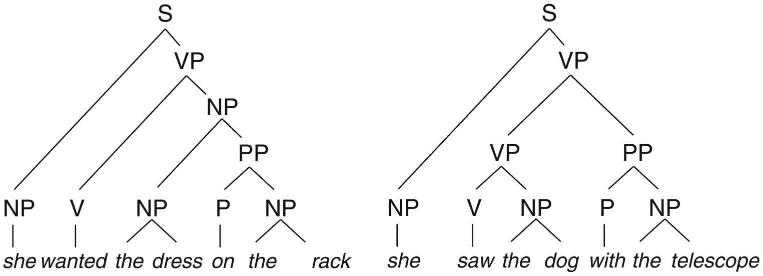

S
VP
NP
PP
NP
V
NP
P
NP
she wanted the dress on the rack
S
VP
VP
PP
NP
V
NP
P
NP
she saw the dog with the telescope

Fig. 4 A small training set of two tree structures (original Figure in [13])

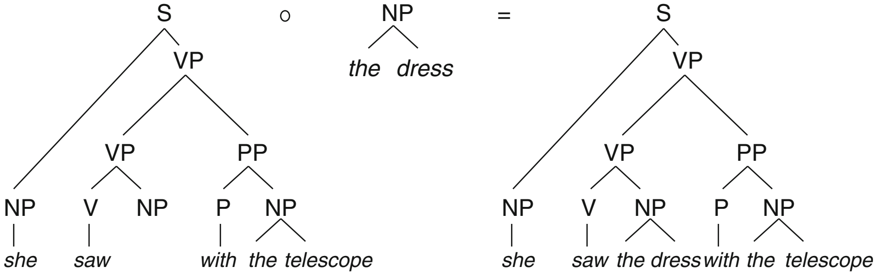

S
VP
VP
PP
NP
V
NP
P
NP
she saw with the telescope
o
NP
the dress
=
S
VP
VP
PP
NP
V
NP
P
NP
she saw the dress with the telescope

Fig. 5 The DOP approach of analysing a new sentence by combining subtrees (original Figure in [13])

available for each of the sentences (see Fig. 4 for two examples). These example derivations are used as a training set by the DOP system, after which the system starts to parse new sentences, i.e. it derives the most plausible structures for the new sentences provided to the system. These newly parsed sentence structures result from a recombination of parts of previously acquired sentence structures (see Fig. 5). DOP has also been applied to problem-solving and equational reasoning in [56]. Thus first experiments with applying DOP to reasoning have been made.

Assuming that parsing a new sentence with the DOP system can also be analysed using the linkography method, which is usually not done, we hypothesise that this would result in a linkograph as in Fig. 6. In this case, each recombination of parts of previously acquired sentence structures, as displayed in Fig. 5, would presumably coincide with a number of recurring links from the new sentence to the sentence structures used for the recombination (Fig. 6). As such, the classical DOP system appears to simulate a kind of analogical reasoning process similar to the reasoning process involved in architectural design and construction, in the sense that records of previous experiences are mapped onto the new ‘target’ experience in order to acquire its (most probable and economic) meaning and proceed in the right direction.

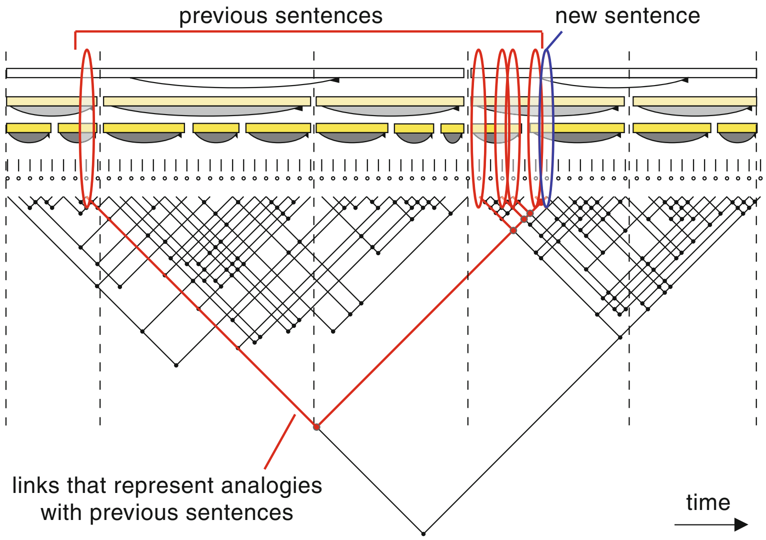


Fig. 6 In interpreting the structure of a new sentence, the DOP approach recombines parts of previously interpreted sentences (see also Fig. 5), thereby relying on an analogical reasoning process between the current sentence and known sentence structures

Using the traditional DOP approach for simulating model-based reasoning in architectural design and construction would presumably result in the following process:

1. A designer is asked at a specific moment in the timeline to annotate a large set of 'design products' (sketches, conversations, buildings, information models, and so forth).
2. The resulting corpus is considered as representative for the complete timeline of episodes experienced by this designer before that specific moment.
3. The resulting corpus is used as a training set by the system.
4. It is tested to what extent the output of the system for new design situations resembles the output of the designer.

The above process has one main disadvantage. Namely, the corpus that is being used for training the system is actually not representative for the complete timeline of the person before the moment of annotation. First of all, only some of the 'products' are considered, whether they be sentences or sketches, thereby leaving out all other possibly relevant products in the actual timeline of the considered person. As a result, the training set represents a timeline with gaps, as illustrated in Fig. 7, in which possibly relevant episodes are missing. Considering that a designer might rely on virtually any of the previous episodes in the analogical reasoning process, this is a significant shortcoming in the system. Additionally, what is included in the system, are not the actual episodes as they were structured at the moment when the person experienced them, they are *remembrances* of

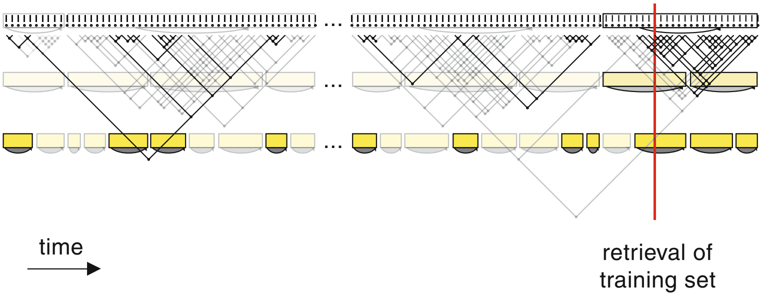


Fig. 7 The linkograph timeline, with an indication of the moment at which an annotated corpus is created by a certain human annotator, and of the number of design episodes that are taken into account for building the annotated corpus

previous episodes as they were structured at the moment when the person had to annotate them. This is a small, but important difference, making the training set even less representative for the future parsing tasks. This issue might well be considered surmountable in the context of textual parsing, it is a high burden in a context of architectural design and construction.

We argue here that the training set needs to take into account more episodes of the linkograph timeline, preferably the complete timeline, to enable a useful level of 'parsing' architectural design knowledge from architectural design 'products'. This objective might be addressed with the unsupervised DOP (U-DOP) approach that is suggested and documented in [13]. This approach namely lets a parser start from scratch and lets it learn in an unsupervised manner from non-annotated sentences. If one would want to apply this approach to architectural design and construction, the main question would be how to structure the diverse kinds of incoming data. The incoming data, namely, do not only include textual data, as is the case in the current U-DOP system [13], but also sketches, dialogues, and so forth. In order to find an initial answer to this research question, we will briefly revisit one of the foundations for model-based reasoning, namely Peirce's process of inquiry, and investigate to what extent it can help us build a U-DOP system for architectural design and construction.

4 Revisiting the Basics: Peirce's Process of Scientific Inquiry

4.1 Revisiting Peirce's Process of Scientific Inquiry

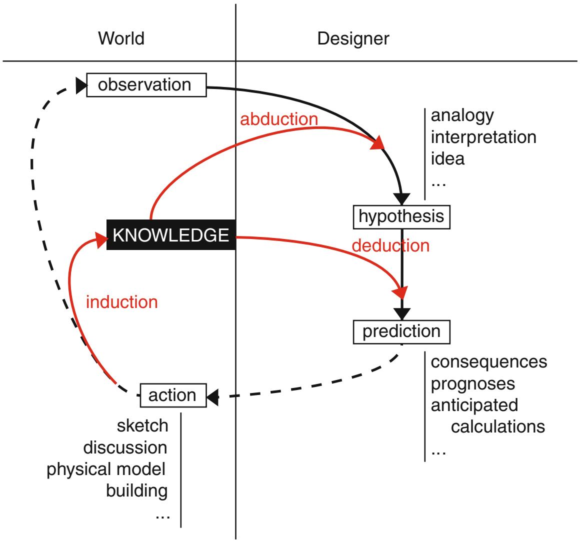
An initial outline of our investigation of Peirce's process of inquiry and its different interpretations can be found in [16], including some resulting conclusions about the usage of information systems in architectural design and construction.

Based on these explorations, we constructed our own interpretation of Peirce's process of inquiry in the context of architectural design and construction (Fig. 8). Note again that this interpretation is first and foremost meant to capture the cognitive processes involved in producing an architectural design or a building. In further refining the details of this interpretation, the role of abductive reasoning might be reconsidered, because we are dealing here with (technological) innovation or invention rather than with discovery (see also [37, 38]).

Central in our interpretation stands a cycle of abductive, deductive and inductive reasoning which is iterated continuously with the surrounding world as its subject. By making this specific combination of all three reasoning modes, the human mind is supposedly able to make hypotheses and interpretations of incoming information, to make predictions and calculations, to devise experiments and learn, all based on the experiences one continuously goes through. The experiential knowledge built up from the continuous iteration through the cycle that is shown in Fig. 8 is considerably different from the far more static kind of knowledge typically embedded in currently available information systems. It can to some extent be considered similar to the type of knowledge that is currently considered by the non-omniscient agent system and the U-DOP system that were both mentioned earlier in this paper. The 'dynamic knowledge' in these systems similarly contrasts with the more static versions in EL systems and in the traditional DOP systems. Of course, an even better and more obvious comparison can be made to model-based reasoning itself.

In the context of a sketching episode, one could explain the diagram in Fig. 8 as following. The situation consists of the reasoning agent or *designer* that is sketching, and the surrounding environment or *world* of this agent. The reasoning agents starts in the top left of the diagram in Fig. 8, namely with an *observation*. In

Fig. 8 Our interpretation of Peirce's process of inquiry in the context of architectural design [16]



this case, this observation most likely consists of the agent seeing a table, a paper, a pencil, and a number of lines that he has previously been drawing on the paper. This observation is to some degree surprising. From this observation, the agent starts an *abductive* reasoning line, which is likely to be rather intuitive in this example. The purpose of this abductive reasoning line is to find an explanation for the agent's observation. We call this explanation also *analogy*, *interpretation* or *idea* in Fig. 8. In this example, this step might, for instance, result in the interpretation that the diverse lines on the paper represent a residential housing unit in floor plan, in which the configuration of the kitchen is not yet ideal, in the sense that a specific other configuration might improve the design. From this *hypothesis*, the reasoning agent makes a specific *prediction* about what to *do* next, resulting in *consequences*, *prognoses*, and/or *anticipated calculations*. Note that the prediction follows quite naturally from the hypothesis, and is typically considered to be a rather straight-forward *deductive* reasoning step. In this example, the reasoning agent might predict that the kitchen can be reconfigured by specific rearrangements of some of the sketch lines, and that, by doing so, the design will improve. The resulting prediction is then put into practice in one way or another (*sketch*, *discussion*, *physical model*, *building*, and so forth). In this example, the reasoning agent alters the sketch, and tries to verify to what extent the kitchen is reconfigured and the design has improved. Depending on this verification, the reasoning agent learns by *induction*. This inductive step is understood in Peirce's definition: *"The purpose of Deduction, that of collecting consequents of the hypothesis, having been sufficiently carried out, the inquiry enters upon its Third Stage, that of ascertaining how far those consequents accord with Experience, and of judging accordingly whether the hypothesis is sensibly correct, or requires some inessential modification, or must be entirely rejected. Its characteristic way of reasoning is Induction"* ([32], p. 6.469).

We argue here that this process should not (only) be understood in the sense of grand steps of discovery in a major architectural design and construction project. Namely, if Peirce's process of inquiry is to be placed in the linkograph timeline that we suggested earlier (see Figs. 1, 2, 3, 4, 5, 6, 7), the process of inquiry is not only in play in the largest design episodes (up in the episodes hierarchy for the linkograph timelines), which might represent complete design projects, but also in the smallest of design episodes (down in the episodes hierarchy for the linkograph timelines), which might represent 'simple' sketches or 'simple' dialogues. This suggested layering of inquiry processes in combination with the linkograph principle is schematised in Fig. 9.

The earlier distinction between within-episode links and cross-episode links can be placed in a new light when considering the suggested layering of inquiry processes. In this case, namely, both link types can be considered identical, but only taking place in a process of inquiry on a different level. Second, this layered process of inquiry might enable to explain why abductive reasoning is so often recognised in very diverse contexts, including on the one hand discovery processes in which complex and high-level abductively obtained ideas are central, and including on the other hand also the notion of abduction in typically small-scale

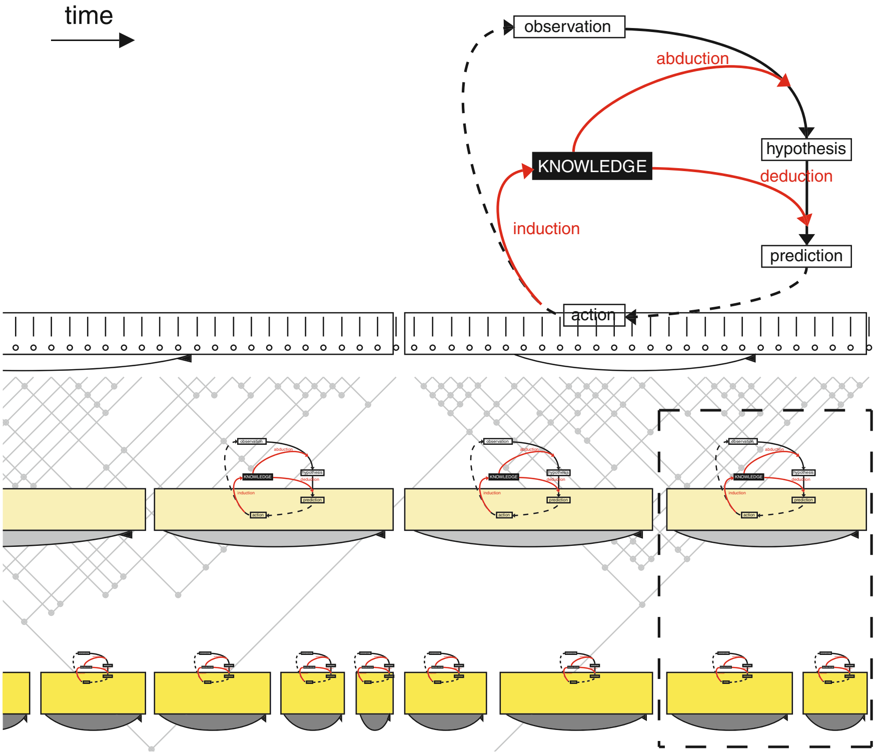


Fig. 9 An indication of the diverse levels in which the process of inquiry, as represented in Fig. 8, can be distinguished in a linkograph sequence

and low-level instinct or intuition (see, for instance, [57] for a discussion about abduction as instinct as opposed to abduction as inference).

Assuming that the process of inquiry takes place at different levels of cognition (Fig. 9), the development of knowledge and beliefs, which is central in the schema in Fig. 8, similarly takes place at different levels of cognition. This situation gives an idea why a logical DEL system with non-omniscient agents, as documented in [14], is so important. Such a system, namely, enables to represent the *sequential changes* in beliefs and knowledge of the logical agents that go through a process as schematised in Fig. 9 at diverse cognitive levels.

4.2 Simulating Peirce's Reasoning Cycle and Building an Autonomous Reasoning Agent

In the remainder of this paper, we document a part of our efforts (1) in simulating the reasoning cycle outlined in Fig. 8 using currently available technologies, and

(2) outline how this simulation might be used for building an autonomous reasoning agent that goes through a process similar to the one depicted in Fig. 9, more precisely the part that is marked in the bottom right corner. These efforts have resulted in an environment in which a reasoning agent processes basic colour information and gradually builds up interpretations of such colours based on incoming information. At this stage of research, such colour information is considered representative for the kind of information typically available in our environment or world. The main hypothesis concluded from this experimental environment states that Peirce's reasoning cycle might be configured in reasoning levels, with each level handling information or patterns in different levels of invariance and meaning.

4.3 Outset of the Experiment

In simulating the process in Fig. 8, we rely on standard semantic web technologies [58, 59], such as Notation3 (N3) [60] and N3Logic [61], which enable the explicit representation of data and rules in a well-defined generic format, and the EYE reasoning engine [62], which enables reasoning processes with this data and rules. The experiment handles a simple colour recognition case. In this case, a reasoning agent is confronted with a random sequence of diverse RGB colour codes and is supposed to interpret this colour code and name it with a colour name. Based on personal background experiences, it is possible that people assign different colour names to identical colours. We therefore consider this colour recognition case as a good example of a simple interpretative process.

The reasoning agent in our experiment has access to a set of previous colour observations or 'colour experiences' that grows with newly encountered colour experiences. The learning process is thus similar to the learning process in the U-DOP approach. The initial set of experiences currently consists of 256 × 256 × 256 RGB colour codes and corresponding colour names. These are described in disjoint graphs, as displayed in Fig. 10.

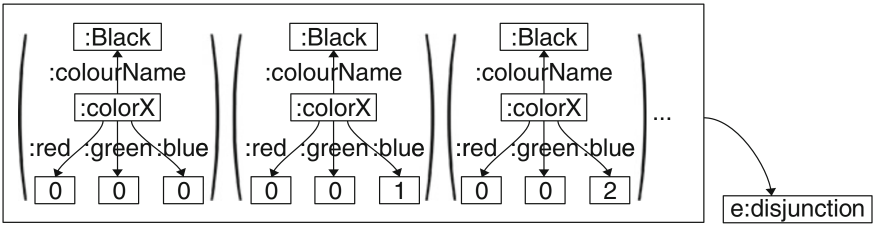


Fig. 10 The reasoning agent has access to a set of disjoint experiences, with each experience making a connection between a specific RGB colour code and a corresponding colour name

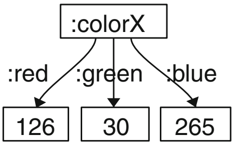


Fig. 11 Representation of one of the colours in the random sequence of colour observations that is interpreted by the reasoning agent

4.4 The Reasoning Process of the Reasoning Agent

In each reasoning cycle, the reasoning agent is confronted with the next colour in a random sequence of colours. This colour is only described by an RGB code and thus not with a colour name (Fig. 11). Based on its set of experiences, the reasoning agent starts an abductive reasoning process that retrieves the most likely colour name of the RGB code at hand. More precisely, the newly encountered RGB code (Fig. 11) is compared to each of the colour codes in its available set of experiences (Fig. 10) to find out whether or not a specific colour name can be a valid 'interpretation' of the RGB code. From this overall process, the reasoning agent calculates which is the most probable interpretation for the RGB code at hand.

According to the process illustrated in Fig. 8, the resulting hypothetical interpretation is now used in a deductive reasoning process to infer predictions. The autonomous reasoning agent should then act according to this prediction and check the validity of the abductively obtained hypothesis in an inductive validation step. One might implement the deductive reasoning step in diverse ways. However, the main idea is that, by temporarily following a certain hypothesis, consequences can be inferred with some degree of belief *from available background knowledge*. Fact is thus that the deductive reasoning step involves extra background knowledge, which is unavailable in the experiment as it is explained above. From the hypothesis that a certain colour code is 'blue' and the currently available background knowledge, one cannot infer a prediction. In the remainder of this paper, we therefore continue with a thought experiment focusing on where to get the required extra background knowledge from and how to combine this with the idea of linkographs as it is depicted in Fig. 9.

4.5 Layered Reasoning Cycles

As suggested earlier in this paper, reasoning agents, like designers, proceed in a sequential manner through (design) episodes. Each episode hereby typically combines a number of actions that have close affinity to the same purpose. For instance, one design episode can be the sketch process of one building part. Additionally, these design episodes form a layered structure. Suppose that each episode can be described with one iteration through the cycle of abductive, deductive and inductive reasoning. In this case, the episode for one part of the building starts from one hypothesis, in this case the assumption that this building part can be improved in one way or another. This episode might fit in a bigger episode, which might in this case be ‘the preliminary design phase’, ‘the phase in which a cheaper version needs to be found of a preceding design’, or anything similar depending on the situation. In this understanding, all the sketch episodes, which are processes of inquiry on their own, actually fit in a larger process of inquiry, which steers the sketch episodes in a specific direction. The sketch episode for a certain part of a building might in turn consist of a set of smaller design episodes, in which even smaller processes of inquiry are at play: the horizontal line episode, the ellipse episode, the hatch episode, and so forth. None of these episodes is ever the same, but in each episode an analogy is made to a previous episode.

Such an upper-level steering of the process of inquiry might be considered for our colour recognition experiment. In this case, it seems reasonable to assume that the colour, for which a name is hypothesised, is part of a broader context and that the deductive reasoning step takes this broader context into account. For instance, when looking at the uppermost light of a traffic light, one typically predicts seeing the colour ‘red’, because this is part of the known pattern ‘traffic light’. Looking at a point with an RGB code that is interpreted as ‘red’ might further confirm that one is looking at the red light of a traffic light. One thus appears to start from high-level patterns, which are decomposed in testable low-level patterns. A similar interaction between low-level, ‘specific’ patterns and high-level, ‘invariant’ patterns is suggested in [63] as an explanation for intelligence and observation in the human mind. The notion of layers in abductive reasoning is also briefly mentioned in [37], although limited to the notion of layered hypotheses.

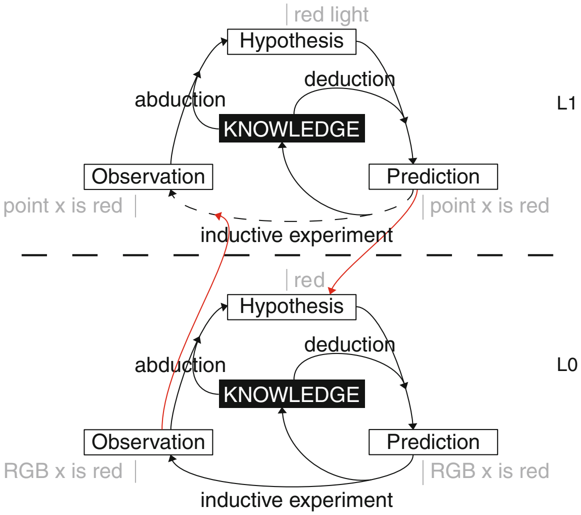


Fig. 12 Possible layering of reasoning cycles. This figure is a detailed view that we suggest for the *bottom-right* portion of Fig. 9

The hypothesis that the reasoning cycle outlined by Peirce runs in different levels could explain how a reasoning agent might build up meaning and information through the construction of high-level patterns from low-level data. The main question is now how these reasoning cycles are connected to each other. We give an indication in Fig. 12 of how this might occur for two reasoning levels, using the context of the experiment started above.

The reasoning agent in the upper level (L1—Fig. 12) handles colour patterns (red-blue-blue-green-blue-orange for instance), whereas the lower level (L0) handles RGB patterns (20-149-21 for instance). Following the traffic light example, one might hypothesise that the next thing to recognise, is the pattern ‘red light’ (Fig. 12, top). From a combination of this hypothesis and the background knowledge or set of known patterns, the deductive reasoning step infers the most economic and valuable prediction. This might in this case be that a specific number of observed points within the expected circle should be categorised as ‘red’. Because this prediction does not include any testable physical data, this prediction is passed on to the reasoning agent in L0 which serves as a surrogate for the inductive experiment.

The L0 reasoning agent supposedly considers the given prediction as a new hypothesis, and infers deductively that the RGB code of the next point should be categorised as ‘red’ for the hypothesis to be true. This can be tested in the physical world, so an inductive experiment is started. In this experiment, the RGB code of a specific point is considered, and the L0 reasoning agent checks whether the predicted colour category matches the considered RGB code. If confirmed, the outcome is stored as a new experience in the background knowledge of the lower-level reasoning agent. This confirms also the inductive experiment of the L1 reasoning agent, leading similarly to extra background knowledge for this reasoning agent. The process continuously starts anew with new actions and new observations.

4.6 Results and Conclusions of the Experiment

The (thought) experiment above gives an idea of how Peirce’s process of inquiry might be configured in layers. Extending this idea in the upper direction in a bottom-up approach might supposedly and eventually result in a reasoning system that is able to work with architectural design and construction information in a way similar to the way in which we work with this information. Before one can work in the upper direction, however, it might be necessary to go further down first, and see how the most basic information can be processed with the suggested setting. When looking at the linkographs that were displayed throughout this paper and the layering principle that was formulated in Fig. 12, one might wonder just how small a (design) episode really is? In other words, up to which atomic levels is a process of inquiry still in play? Does one need to look into audio frequencies and levels of light intensity? Does one need to go into microscopic detail? Or is one of the higher levels the place to stop? We provisionally suggest here to look at the level of information that we are able to consider as human beings.

A second conclusion that we draw from our initial exploratory experiment concerns the labelling of patterns and thus of 'knowledge'. In the colour recognition experiment, we represented observations of colours using RGB codes together with annotated colour names. The colour codes might be representative of the way in which we perceive colours through measures of light intensity in the three visible spectra red, green and blue. The colour names that are assigned to the diverse colour codes, however, certainly are *not* how we remember colours. Many more elements are of relevance in our appreciation of a certain colour. Following the idea of layered reasoning cycles, one might also consider a layered structure in our understanding of concepts. With layer height, these layers become increasingly invariant and complex, mainly because this layer does not need to know the precise details of the concept it is dealing with. So, the only thing that the reasoning system theoretically needs to know are the empirical details at the lowest layer, and all the upper-level layers are but patterns that refer to these empirical details (thereby bringing a personal and experience-based structure to those empirical details). So, in terms of the colour recognition experiment, the only *data* that should be stored are the basic patterns of RGB codes. In the case of auditory data, only the sequence of sound frequencies should be stored as *data*. Besides that, only patterns of those data should be maintained and used, according to theory.

5 Conclusion

We have looked in this paper into the kind of knowledge that is handled during architectural design and construction projects by individual designers. We have hereby relied on existing theories of design thinking. In particular, we looked into the possible parallels between these theories and model-based reasoning. The resulting overview in the beginning of this paper outlines how architectural designers and experts in construction industry typically go through a rather pragmatic process, in which they (1) continuously try to make sense of their design situation using their own background knowledge (abductive reasoning), (2) subsequently make predictions about how to react upon this design situation (deductive reasoning), and (3) finally learn from the actions they make by revising their background knowledge (inductive reasoning).

Second, we have indicated in this paper how the model-based reasoning features of architectural design and construction projects can be recognised in explicit representations of the architectural design process, in this case linkograph representations. From a linkograph representation, one can see how designers proceed in episodes and make both within-episode links (in the pursuit of a specific idea) and cross-episode links (analogies to previous episodes). This has resulted in the notion of reasoning layers or levels: diverse episodes might be combined into a greater episode on a higher level, thereby transforming cross-episode links into within-episode links in this higher level.

We have subsequently looked into two recent conceptual models or simulations of reasoning processes similar to model-based reasoning processes, namely non-omniscient agents in a DEL environment, and the DOP parsing system. Both systems give a good indication of how model-based reasoning processes might be implemented or simulated. An implementation of a reasoning system that can handle the kind of knowledge that is typically handled in architectural design and construction projects, however, requires far more detail and a notably higher level of complexity than can currently be handled, for instance in the classical DOP approach. This level of complexity might be realised by an appropriate implementation of the layers that were identified in the linkograph representations. In order to find out how we can build a simulated environment with such features, we finally looked into prevailing interpretations of Peirce's process of inquiry (cycle of abductive, deductive and inductive reasoning), eventually suggesting a layered configuration of this process of inquiry aligned to the linkograph idea. Note again that the exact role of abductive reasoning for architectural design is subject to further research, because this type of reasoning is typically considered only in the context of discovery, which includes a search for explanatory hypotheses in the event of surprising events. Architectural design and construction, on the other hand, is typically understood as an ill-structured problem solving process, in which a search is needed for actions in design situations that display a sort of need. An initial implementation of the Peirce's process of inquiry was nevertheless proposed and an initial and exploratory experiment was documented for a colour recognition case.

Although the colour recognition experiment is first and foremost a thought experiment, it does give some idea of how Peirce's process of inquiry might be configured in layers and how such a configuration might eventually result in the simulation of reasoning processes in specific practices, including architectural design and construction. However, key in building a system that autonomously goes through a reasoning process is the way in which the agents on the diverse reasoning levels or layers interact. As it is suggested now, the prediction that is generated by the upper reasoning agent (L1) supposedly results in the hypothesis of the lower reasoning agent (L0). This implies that the lower reasoning agent might not actually go through an abductive reasoning phase itself. However, this is only one of the possible configurations, and a lot more research and testing in diverse more complex contexts is necessary for further confirmation or refutation of this configuration, and of the very idea of layering Peirce's process of inquiry.

Acknowledgments The authors would like to thank the reviewers, both within and outside the official reviewing process, for their valuable comments and suggestions to improve the paper. We are confident that these comments and suggestions will be of value also to future research.

References

1. Magnani, L.: Abductive reasoning: philosophical and educational perspectives in medicine. In: Evans, D., Patel, V. (eds.) Advanced Models of Cognition for Medical Training and Practice, pp. 21–41. Springer, Berlin (1992)
2. Jovanovic, A., Krneta, G.: Abductive reasoning and second language learning. J. Lang. Teach. Res. **3**(2), 306–313 (2012)
3. Wirth, U.: Abductive reasoning in Peirce's and Davidson's account of interpretation. T. C. S. Peirce Soc. **35**(1), 115–127 (1999)
4. Arrighi, C., Ferrario, R.: Abductive reasoning, interpretation and collaborative processes. Found. Sci. **13**, 75–87 (2008)
5. Shelley, C.: Visual abductive reasoning in archaeology. Philos. Sci. **63**(2), 278–301 (1996)
6. Thagard, P.: The cognitive-affective structure of political ideologies. In: Martinovski, B. (ed.) Emotion in Group Decision and Negotiation. Springer, Berlin (forthcoming)
7. Shelley, C.: Motivation-biased design. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 2956–2961 (2011)
8. Thagard, P.: Creative combination of representations: scientific discovery and technological invention. In: Proctor, R., Capaldi, E. (eds.) Psychology of Science: Implicit and Explicit Processes, pp. 389–405. Oxford University Press, Oxford (2012)
9. Goldschmidt, G.: Linkography: assessing design productivity. In: Trappl, R. (ed.) Cybernetics and Systems '90. World Scientific, Singapore (1990)
10. Goldschmidt, G.: Criteria for design evaluation: a process-oriented paradigm. In: Kalay, Y. (ed.) Evaluating and Predicting Design Performance. Wiley, New York (1992)
11. Goldschmidt, G.: The designer as a team of one. Des. Issues **16**(2), 189–209 (1995)
12. Bod, R.: The data-oriented parsing approach: theory and application. In: Fulcher, J., Jain, L. (eds.) Computational Intelligence: A Compendium, pp. 307–342. Springer, Oxford (2008)
13. Bod, R.: From exemplar to grammar: a probabilistic analogy-based model of language learning. Cog. Sci. **33**, 752–793 (2009)
14. Velazquez-Quesada, F.: Small steps in dynamics of information. Ph.D. thesis, University of Amsterdam, Amsterdam (2011)
15. Pauwels, P., Bod, R.: Including the power of interpretation through a simulation of Peirce's process of inquiry. Lit. Ling. Comput. (in press). doi: 10.1093/llc/fqs056
16. Pauwels, P., De Meyer, R., Van Campenhout, J.: Design thinking support: information systems vs. reasoning. Des. Issues **29**(2) (2013)
17. Archer, L.: Systematic methods for designers. In: Developments in Design, Methodology, pp. 57–82. Wiley, Chichester (1965)
18. Jones, J.: Design Methods: Seeds of Human Futures, 1st edn. Wiley, New York (1970)
19. Alexander, C.: The state of the art in design methods. Des. Methods Group Newslett. **5**(3), 3–7 (1971)
20. Jones, J.: How my thoughts about design methods have changed during the years. Des. Methods Theor. **11**(1), 48–62 (1977)
21. Rittel, H., Webber, M.: Planning problems are wicked problems. In: Cross, N. (ed.) Developments in Design Methodology, pp. 135–144. Wiley, Chichester (1984)
22. Simon, H.: The structure of ill-structured problems. Artif. Intell. **4**, 181–201 (1973)
23. Rittel, H., Webber, M.: Dilemmas in a general theory of planning. Policy Sci. **4**, 155–169 (1973)
24. Schön, D.: The Reflective Practitioner: How Professionals Think in Action. Temple Smith, London (1983)
25. Lawson, B.: How Designers Think: The Design Process Demystified, 4th edn. Elsevier, Oxford (2005)
26. Cross, N.: Designerly Ways of Knowing. Springer, London (2006)
27. Goldschmidt, G.: The dialectics of sketching. Des. Stud. **4**, 123–143 (1991)

28. Goldschmidt, G.: On visual design thinking: the vis kids of architecture. Des. Stud. **15**(2), 158–174 (1994)
29. Cross, N.: Designerly ways of knowing. Des. Stud. **3**(4), 221–227 (1982)
30. Douglas, M., Isherwood, B.: The World of Goods. Allen Lane, London (1979)
31. Cross, N.: The nature and nurture of design ability. Des. Stud. **11**(3), 127–140 (1990)
32. Peirce, C.: Collected Papers of Charles Sanders Peirce. vols. 1–6 (Eds. C. Hartshorne & P. Weiss) (1931–1935), vols. 7–8 (Ed. A.W. Burks) (1958). Harvard University Press, Cambridge (1958)
33. March, L.: The logic of design and the question of value. In: The Architecture of Form, pp. 1–40. Cambridge University Press, Cambridge (1976)
34. Bogen, J.: The other side of the brain II: an appositional mind. Bull. Los Angeles Neurol. Soc. **34**(3), 135–162 (1969)
35. Simon, H.: Models of Discovery and Other Topics in the Methods of Science. Reidel, Dordrecht (1977)
36. Saunders, D., Thagard, P.: Creativity in computer science. In: Kaufman, J., Baer, J. (eds.) Creativity Across Domains: Faces of the Muse. Lawrence Erlbaum Associates, Mahwah (2005)
37. Thagard, P., Millgram, E.: Inference to the best plan: a coherence theory of decision. In: Ram, A., Leake, D. (eds.) Goal-Driven Learning, pp. 439–454. MIT Press, Cambridge (1997)
38. Thagard, P., Croft, D.: Scientific discovery and technological innovation: ulcers, dinosaur extinction, and the programming language java. In: Magnani, L., Nersessian, N., Thagard, P. (eds.) Model-Based Reasoning in Scientific Discovery, pp. 125–137. Kluwer Academic/Plenum Publishers, New York (1999)
39. Ennis, C., Gyeszly, S.: Protocol analysis of the engineering systems design process. Res. Eng. Des. **3**(1), 15–22 (1991)
40. Cross, N.: Design cognition: results from protocol and other empirical studies of design activity. In: Eastman, C., McCracken, W., Newstetter, W. (eds.) Design Knowing and Learning: Cognition in Design Education, pp. 79–104. Elsevier, Oxford (2001)
41. Kavakli, M., Gero, J.: The structure of concurrent cognitive actions: a case study of novice and expert designers. Des. Stud. **23**(1), 25–40 (2002)
42. Kolko, J.: Abductive thinking and sensemaking: the drivers of design synthesis. Des. Issues **26**, 15–28 (2010)
43. Ericsson, K., Simon, H.: Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge (1993)
44. van Someren, M., Barnard, Y., Sandberg, J.: The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes. Academic Press, San Diego (1994)
45. Kan, J., Gero, J.: Acquiring information from linkography in protocol studies of designing. Des. Stud. **29**(4), 315–337 (2008)
46. Gentner, D., Bowdle, B., Wolff, P., Boronat, C.: Metaphor is like analogy. In: Gentner, D., Holyoak, K., Kokinov, B. (eds.) The Analogical Mind: Perspectives from Cognitive Science. MIT Press, Cambridge (2001)
47. Grace, K., Saunders, R., Gero, J.: Interpretation-driven visual association. In: Proceedings of the Second International Conference on Computational Creativity, pp. 132–134 (2011)
48. Lakoff, G., Johnson, M.: The metaphorical structure of the human conceptual system. Cogn. Sci. **4**(2), 195–208 (1980)
49. Hofstadter, D.: Analogy as the core of cognition. In: Gentner, D., Holyoak, K., Kokinov, B. (eds.) The Analogical Mind: Perspectives from Cognitive Science. MIT Press, Cambridge (2001)
50. Ward, T.: Analogical distance and purpose in creative thought: mental leaps versus mental hops. In: Holyoak, K., Gentner, D., Kokinov, B. (eds.) Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences. NBU Series in Cognitive Science. NBU Press, Sofia (1998)
51. Heylighen, A.: Building memories. Build. Res. Inf. **35**, 90–100 (2007)
52. Cross, N.: Natural intelligence in design. Des. Stud. **20**(1), 25–39 (1999)

53. Jones, J.: Design methods reviewed. In: Gregory, S. (ed.) The Design Method. Butterworth, London (1966)
54. van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Synthese Library Series, vol. 337. Springer, Heidelberg (2007)
55. van Benthem, J.: Logical Dynamics of Information and Interaction. Cambridge University Press, Cambridge (2011)
56. Bod, R.: Getting rid of derivational redundancy or how to solve Kuhn's problem. Mind. Mach. **17**(1), 47–66 (2007)
57. Paavola, S.: Peircean abduction: instinct or inference? Semiotica **153**, 131–154 (2005)
58. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 35–43 (2001)
59. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. Int. J. Semant. Web Inf. **5**(3), 1–22 (2009)
60. Berners-Lee, T., Connolly, D.: Notation 3 (N3): a readable RDF syntax. W3C team submission. http://www.w3.org/TeamSubmission/n3/
61. Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., Hendler, J.: N3Logic: a logical framework for the world wide web. Theor. Pract. Log. Prog. **8**(3), 249–269 (2008)
62. De Roo, J.: Euler proof mechanism. http://www.agfa.com/w3c/euler/
63. Hawkins, J., Blakeslee, S.: On Intelligence. Times Books, New York (2004)

Models and Ideology in Design

Cameron Shelley

Abstract Research on model-based reasoning in technology tends to focus on the informational role of models. That is, it concentrates on the use of models as a source of information about how to solve a given design problem. However, besides their informational role, models can also serve an ideological role. That is, models can provide affirmation of the moral correctness of a design program. In this article, the role of models within three design programs is examined. These programs are Gothic Revivalism, Modernist architecture, and industrial design in the early twentieth Century. In each case, the ideology of the program is sketched and the role of models within them is discussed. Through this exercise, we see that models for design are selected not only for their help in solving design problems but also for their service in reflecting and reinforcing the worldview of the designer.

1 Introduction

In the original version of the Calculator app for its iPod and iPhone, Apple designed the interface to recall the well-known Braun ET44 calculator created by legendary designer Dieter Rams [1]. The rounded buttons with convex, raised centers were a distinctive design element of the original calculator. Even though the flat, touch-sensitive surface of the iPhone could not accommodate raised buttons, the crisp and candy-like appearance made the interface attractive and even inviting. The point of the imitation was at least two-fold: First, it was a borrowing from a proven design, aimed to make the calculator app straightforward and

C. Shelley (✉)
Centre for Society, Technology, and Values,
University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: cam_shelley@yahoo.ca
URL: http://cstv.uwaterloo.ca

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_33, © Springer-Verlag Berlin Heidelberg 2014

pleasant for users, regardless of their familiarity with the original. Second, it was an homage to a design classic; Jobs's admiration for Rams is well known [2].

Rams's ET44 design likewise stood as a model for the iPhone app in at least two respects, as a source of innovation and as an assurance of value. The term *model* itself reflects this duality of meaning: A model can be source of information, a fund of experience on which designers can call in order to optimize their own designs for a given purpose. A model can also be an exemplar, that is, it can embody intrinsic worth. Consider the difference between a *model solution* to a class assignment and a *model citizen*. The model solution is a solution to some problem that informs students how to approach problems of a particular sort. Used properly, it saves them from the frustration of a long round of designing from scratch. The model citizen is someone who embodies values that the community wishes other citizens to emulate. By recognizing the bravery of someone who has rescued others from a burning building, say, the community is not requiring everyone to do the same. Instead, it is suggesting that bravery and selflessness are values that the community, and everyone in it, should hold in high regard.

Rams's calculator served as a model of design in both senses. It saved Apple designers the trouble of trying to design a calculator interface from first principles. It also provided them with a standard of design excellence that their own work could acknowledge.

Discussion of model-based reasoning in technology has focused on the first, instrumental sense of model, that is, the sense of models as model solutions. There is nothing regrettable in this project as such. However, understanding the role of models in technology cannot be complete without considering how models sometimes acts as model citizens. In this role, models embody values in the ideologies that designers promote. That is, they represent ideals that designers themselves subscribe to, and that they also want or expect their clients to accept.

The aim of this article is to explore the role of models as model citizens in the ideology of technological design. In order to further this exploration, I look at the role of models in three different design movements in modern history.

1. Revivalism: The Gothic Revival of the nineteenth Century looked to the European Middle Ages for models of good design.
2. Modernism: Modernism was a movement in the early and mid 20th Century that rejected historical design and looked instead to contemporary heavy industry for design elements in all types of structures.
3. Industrial design: Industrial designers of the early and mid 20th Century tended to extrapolate into the future in search of elements for contemporary designs.

Each design movement looked to a particular historical period for design models. They did so, in part, for instrumental reasons, that is, for assistance in solving design problems. However, they did so also for ideological reasons, because the models from each historical period represent values that the designers wanted to praise and to promote. In the following sections, the nature of each movement is outlined, and its role in the ideology of contemporary designing is examined. From this examination, the importance of ideology in model selection will become more clear.

2 Revivalism: Looking to the Past

In general, revivalism refers to the use of elements from an historical design style in contemporary designs. One of the best-known revivalist design movements was the Gothic Revival, which reached its height in Britain in the nineteenth Century. The Gothic Revival began in the eighteenth Century, after the antiquarian Horace Walpole wrote the medieval romance, the *Castle of Otranto* in 1764, and another book about his house, Strawberry Hill in 1774. The latter book described how Walpole renovated this house with elements from medieval buildings, such as pointed arches, crockets, quatrefoils, and so on. This book helped to raise general interest in the architecture and design from medieval Europe [3].

In the years after the Napoleonic Wars, the government Church Building Commission undertook to subsidize the construction of hundreds of churches throughout Britain. The general idea was to knit back together the social fabric that had unraveled somewhat under the pressures of the prolonged and ideologically charged conflict. Many of these churches were built in a Gothic style. That is, they often applied design elements from gothic structures, in the spirit of Walpole's version of decoration [4]. See the Gothic Revival St. Peter's church, built under the auspices of the Commission in Fig. 1 for an example.

Some architects criticized the Commission's approach to architecture. Among them was Augustus Welby Northmore Pugin (1812–1852), most famous for his work on the Palace of Westminster in the 1840s. Pugin objected that, oftentimes, Gothic elements such as pointed windows and buttresses were simply tacked on to structures that were essentially classical in design, having the basic form of Greek temples. The resulting, hybrid structures were not truly Gothic at all in his eyes. Compare the neoclassical St. John's church, built under the auspices of the Commission, in Fig. 2 to St. Peter's. Note how St. Peter's imitates the roof profile and basic layout of the neoclassical model, except with Gothic-style buttresses and accoutrements in place of classical ones.

Pugin had formed a strong attachment to medieval buildings and churches in particular. He had converted to Catholicism in 1834, in part as a result of his experiences studying medieval churches in England and northern France. He saw a strong connection between medieval architecture and proper, Christian faith. As his biographer puts it, for Pugin, "the Catholic Church is the true church, Gothic architecture its revealed form, true in the sense of absolute, a divine, revealed form" [5].

For Pugin, the ideology of the Gothic Revival contained at least two values, those being *authenticity* and *conservatism.* Being authentic did not mean, of course, that a building had to be a genuine medieval structure. Instead, it meant that a building should be a close facsimile of genuine buildings of the earlier era.

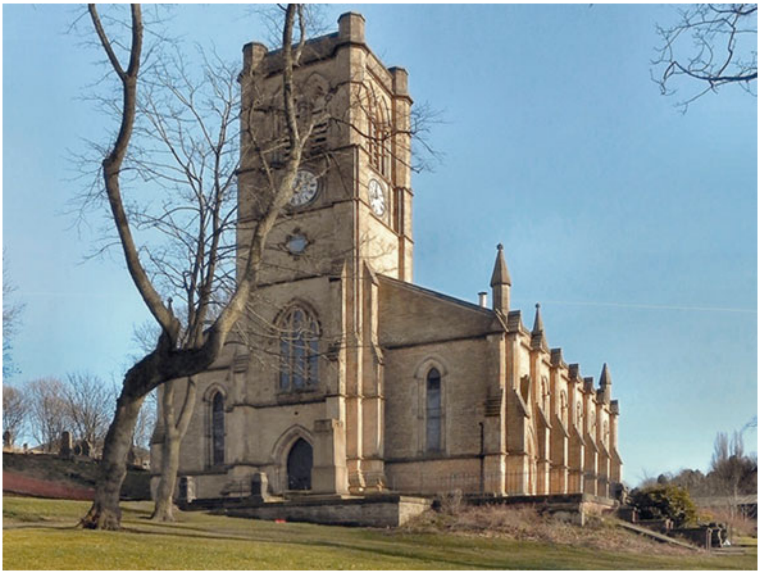


Fig. 1 St. Peter's church, Blackley, UK. Designed by E. H. Shellard, ca. 1845. Photo by David Dixon/Wikimedia commons: http://en.wikipedia.org/wiki/File:St_Peter,_Blackley.jpg

Studies of genuine structures in England and northern France served as models that could guide the revivalist architect in this matter.

An important aspect of authenticity, then, was *localism*. That is, revivalist structures should use local materials and building methods that imitated the methods used in the Middle Ages by builders in the vicinity were the new building was to stand. In the Middle Ages, the transportation infrastructure of northern Europe was neither efficient nor capacious enough to allow for the shipment of large volumes of materials over long distances. As a result, medieval buildings tended to be made of materials acquired in the local area. Similarly, poor infrastructure also meant that the builders hired to construct buildings were recruited from the local area. This situation facilitated the existence of local idioms in design. That is, each region tended to see the rise of design traditions within the region and different from the traditions that arose in other regions. For authenticity in a building to be situated in a given region, Pugin thought it best to observe the local building traditions that characterized genuine, medieval buildings nearby.

From an instrumental standpoint, this emphasis on localism was not always optimal. In the nineteenth Century, it would often be more economical to ship materials and workmen from other areas of the country by rail or canal. Pugin's emphasis on localism was motivated on ideological and not instrumental grounds.

Beyond authenticity, Pugin's careful imitation of medieval design was motivated by religious conservatism. Contemporary Protestant churches tended to be spatially simple in the sense that their interiors were relatively undifferentiated spaces in which the congregation and minister gathered together. A Catholic

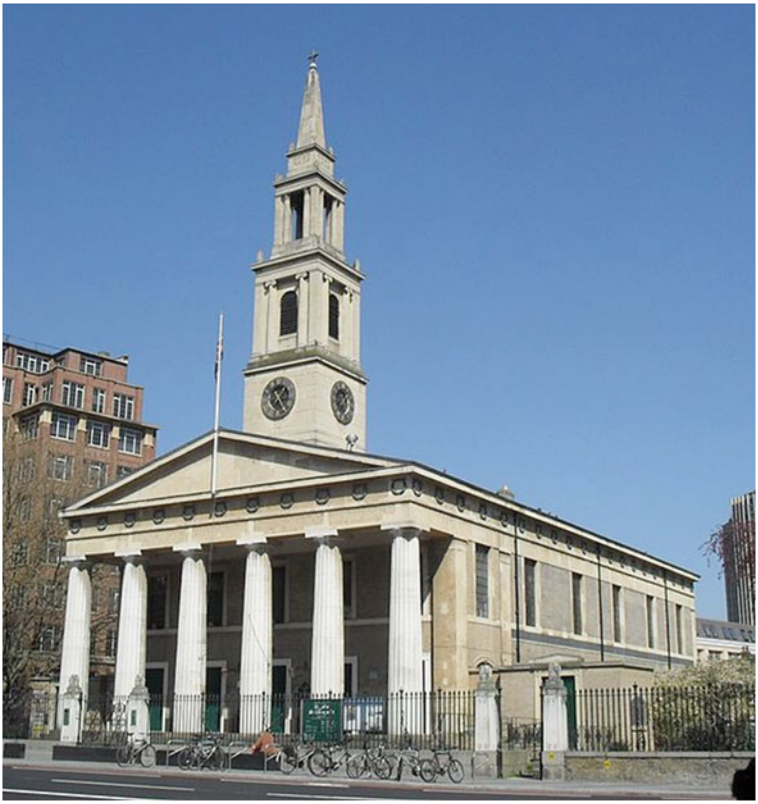


Fig. 2 St. John's church, London, UK. Designed by Francis Octavius Bedford ca. 1824. Photo by Hassocks5489/Wikimedia commons: http://en.wikipedia.org/wiki/File:St_John%27s_Church,_Waterloo_Road,_Waterloo,_London_(IoE_Code_204772).JPG

church from the Middle Ages was a microcosm of the medieval worldview, a hierarchical arrangement of separate spaces, each with its own appropriate functions and occupants. Pugin's church designs persisted in this traditional divided arrangement of spaces. For example, Pugin's church designs usually contained a *rood screen* to separate the nave from the chancel, thus keeping the altar and choir apart from the congregation. This separation had an important function in the medieval ceremony of mass but had fallen out of favor among Anglicans and Catholics by the nineteenth Century. Pugin designed the screens for his churches as a way of advocating for the return to medieval forms of worship, which he regarded as superior. A controversy ensued which was settled, in the end, by an appeal to the Vatican, which sided with Pugin's opponents. As a result, many of the rood screens in his churches were subsequently removed [6].

So, authenticity and religious conservatism were important values in Pugin's version of the Gothic Revival. First, the continued presence of medieval Gothic architecture in the country supported the view that such architecture was an authentic expression of Englishness. Study of regional variations in Medieval Gothic architecture in England only served to reinforce its authenticity. By reviving the Gothic style, then, Pugin was not introducing a foreign element into English life.

Second, Pugin could point to the models as evidence of the Englishness of the Catholic Church. Britain was officially Anglican and had only recently passed a law tolerating Catholicism and allowing Catholics to hold public offices. Pugin hoped to turn this tolerance into broader acceptance of Catholicism, even reconciliation with the Anglican Church. People's attachment to these medieval buildings, reinforced through Pugin's own works, might lead them to reconsider their separation from the Church of Rome.

The Gothic Revival shows how models can play a role in the ideological side of design. Design is undertaken not merely to solve a given problem but to reflect a worldview. In the Gothic Revival, historical authenticity and religious propriety were dominant ideological values that informed design. Practitioners of the Gothic Revival, such as Pugin, naturally looked to surviving instances of Gothic architecture as models of solutions to technical design problems and also as embodiments of their values.

3 Modernism: Transcending History

Historical structures are an obvious source of models within a design ideology that seeks to reinstate or reinvigorate past modes of living. By the outset of the twentieth Century, some designers had decided that revivalism was not a tenable ideology. Technological advances had brought with them new challenges and those challenges called for new approaches. Thus, appropriate design ideology had to be divorced from attachments to the past.

The designer who perhaps best embodies this form of modernism is Le Corbusier. Born as Charles Edouard Jeanneret (1887–1965) in Switzerland, Le Corbusier trained in engraving but made his mark in architecture and urban planning. Le Corbusier set up his practice in Paris towards the end of World War I and advocated modernist housing as a means of quickly rebuilding the housing stock destroyed in the war. He published a collection of essays, *Vers une architecture* [7], discussing his views on architecture.

In this book, Le Corbusier famously promoted his view that the design of contemporary industrial objects formed the best source of models for architecture. In particular, he described a house as "a machine for living", meaning that houses should be designed just like cars, boats, or other mass-produced objects. His most famous house, the Ville Savoye built in 1928–1931 near Paris, was designed in this manner. See Fig. 3. It was made of modern materials, reinforced concrete, steel, and glass, with no attempt at disguise or decoration. The use of steel and concrete structure meant that there was no need for load-bearing walls, so interior walls

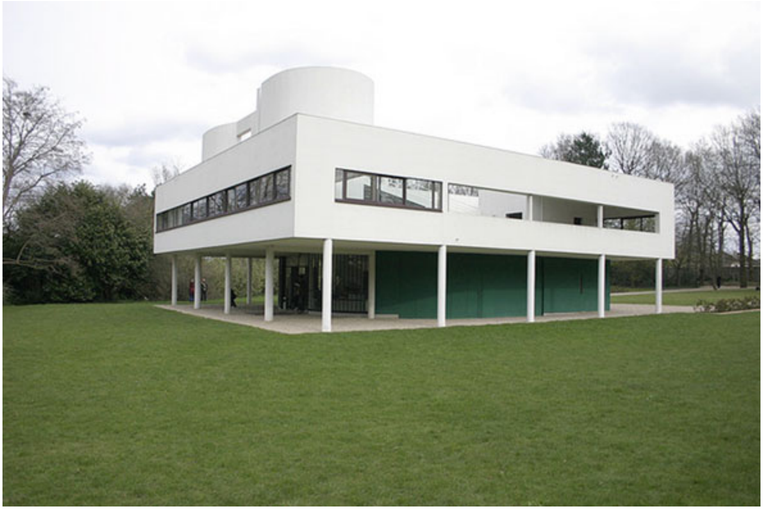


Fig. 3 La Ville Savoye, Poissy, France. Designed by Le Corbusier ca. 1930. Photo by Coronel Rodrigombria/Flickr.com: http://www.flickr.com/photos/coronel_rodrigombia/3313778560/

were eliminated or minimized. Thus, the interior was largely open space. Each of the four exterior walls contained long ribbon windows, reducing the distinction between inside and outside. The roof was flat, not peaked, and contained a garden for the occupants to use, rather like the recreation deck on an ocean-liner. In fact, the design of the house was inspired by the design of the decks of ocean liners such as RMS Aquitania, a picture of which was featured in Le Corbusier's book. See Fig. 4.

In urban planning, Le Corbusier's ideas are conveyed by the *Ville Conptemporaine* exhibit that he mounted at the Salon d'Automne in Paris in 1922 [8]. This ideal city contained a central district consisting of glass skyscrapers and apartment buildings arranged in a grid pattern. Connecting the buildings with each other and the surrounding countryside were a network of highways. At the center where the highways met would be a seven-level transportation hub including levels for railways, highways, subways, with an airport on the top layer. The buildings were to be raised off the ground on stilts so that the ground level of the entire city could be a large greenspace.

Each component of the design is dedicated to the fulfillment of a single function. The buildings in the central core were for white-collar workers who would run the city. Blue-collar workers and industrial facilities would be placed in areas outside the central district. Some buildings in the core were for work, others for housing, others for commerce. Some spaces were dedicated to transportation, e.g., highways and airports whereas others were dedicated to recreation, e.g., rooftops and the ground-floor greenspace. In order to transition from one activity to another, a person would drive a car along the highway to the space designed for that activity.

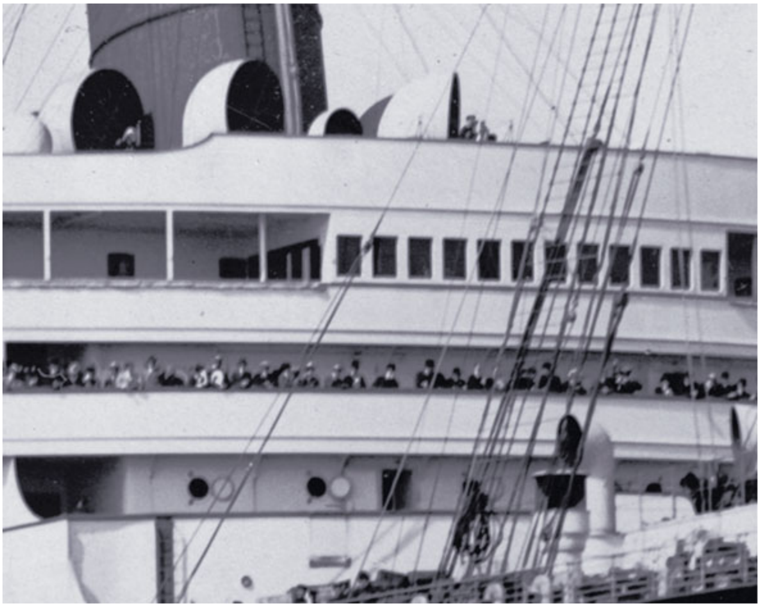


Fig. 4 Deck of the RMS Aquitania. Detail of photo by Library of Congress/Wikimedia commons: http://commons.wikimedia.org/wiki/File:SS_Aquitania.jpg

This initial design was generic. However, Le Corbusier later exhibited a design of this type specifically for Paris at the *Exposition des Arts Decoratifs* in 1925. It was called the *Voisin* plan after the automobile and aircraft manufacturer that sponsored Le Corbusier at the event. See Fig. 5. In this plan, Le Corbusier suggested razing the central district of Paris and replacing it with a grid of eighteen sixty-story skyscrapers connected by highways. Important monuments like Notre Dame would be retained but, otherwise, the core of the City was to be entirely rebuilt. The traditional but chaotic layout of Paris, with its maze of medieval streets, was to be eliminated in favor a functional and efficient grid of massive buildings and roads.

Le Corbusier's ideas become influential particularly after World War II, when building and re-building projects took off all over the industrialized world. Highways were built, old neighborhoods were bulldozed to make way for expressways and apartment blocks, and the downtowns of big cities filled with slabs of concrete, glass, and a steel. Note the resemblance of the Co-op City buildings in Fig. 6 to those envisioned by Le Corbusier.

A great virtue of modernist design is that it made efficient use of industrial materials, allowing infrastructure to be built rapidly and affordably. A great problem of the modernist approach is that it could be overwhelming and inflexible, treating people somewhat as goods to be stored and moved about as required by the design of their infrastructure [9].

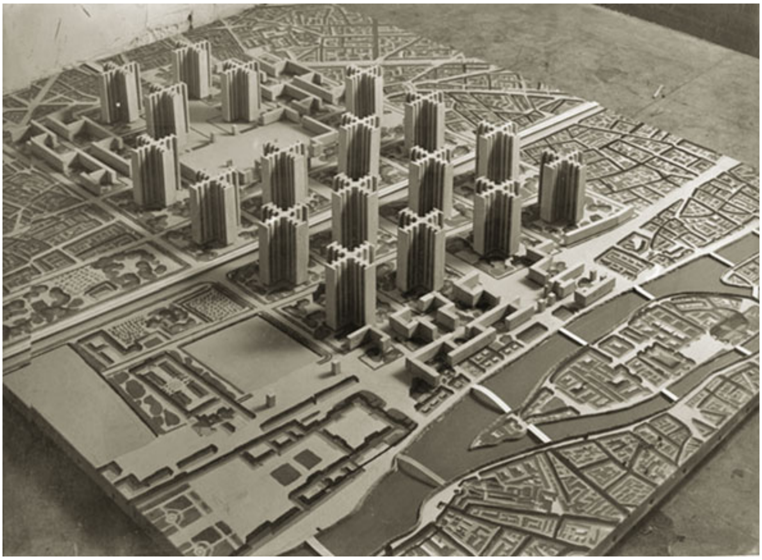


Fig. 5 Model of Voisin plan for Paris, by Le Corbusier, 1925. Photo by Gaynoir/Flickr.com: http://www.flickr.com/photos/gaynoir/4936960279/in/set-72157624723694364/

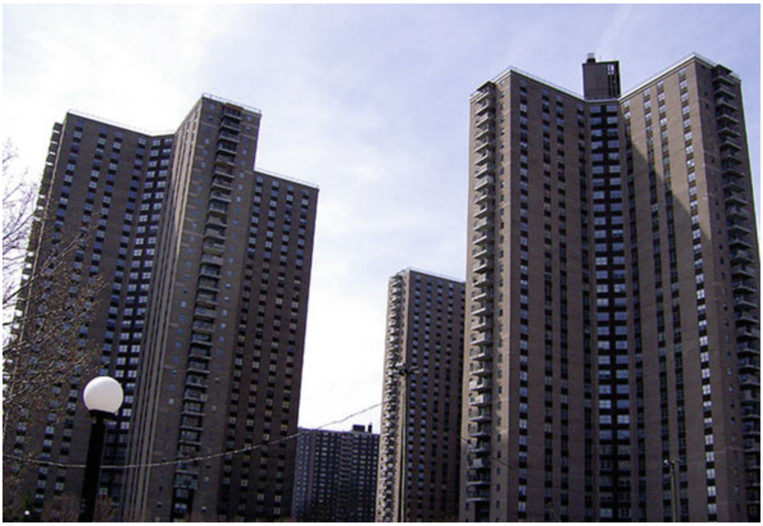


Fig. 6 Co-op City, the Bronx, USA. Note the similarity to the skycrapers in Voisin plan. Photo by Jules Antonio/Flickr.com: http://www.flickr.com/photos/julesantonio/3390575220/

The main values of the modernist view are *functionalism* and *conformity*. Functionalism identifies the analytic approach to design adopted by the modernists, embodied in the expression "form follows function". In brief, design should be approached from perspective independent of history. To a functionalist designer, it is of no relevance how buildings or anything else were designed in antiquity or the Middle Ages. All that matters is the problem to be solved and the means available to solve it with. Le Corbusier took a Platonic view on which the activities of living and the architectural forms for building are regarded as a set of ideal forms. His buildings exhibited a preference for simple geometric figures and solids in combination.

In terms of analysis, the modernist designer sought to disaggregate the various activities that would occur in the use of the design. When designing a house, or machine for living, the process of living should be broken down into its subprocesses and some space designated for each. The remaining task was to place the spaces for each activity in the correct relation to each other. In the Ville Savoye, the ground floor contained spaces for auxiliary functions, including the entrance, the garage, and rooms for chauffeurs and maids. The main floor contained the bedrooms for sleeping, the kitchen for cooking, and the salon for interacting. The roof contained the garden for relaxation and recreation. In the Voisin plan, activities such as resting, recreating, working, and moving were each assigned a separate space. Highways were used to allow people to move from one activity to another with as little hindrance as possible.

Besides disaggregation, functionalism also implies a kind of universalism. Just as the relevance of historical antecedents is minimized, so is the relevance of regionalism. Functionalist design tends to focus on the basics. Living in a house, for example, is analyzed to an almost biological level: eating, sleeping, exercising, interacting. These functions are universal ideals that apply to all people. By contrast, cultural preferences, such as having a porch or enlarged foyer to lend importance to the front entrance, are minor considerations.

In addition, another important component of functionalism is honesty. That is, the aesthetic value of a modernist design should come from its construction instead of from add-on decorations that serve no utilitarian purpose. In the Victorian era, designers might use iron as a structural element of Gothic buildings, as Pugin did for the Palace of Westminster. However, the iron would be hidden from view. From a modernist perspective, this practice is dishonest. A building made of steel, glass and concrete should display its construction. Furthermore, its construction should be such that nothing further is required to make the building look good.

Besides functionalism as such, modernist designs also tend to require *conformity* from their users. A modernist building, for example, does not invite later modification. If the building's form resulted from a correct application of timeless ideals, then no modifications should be necessary. In the Voisin plan, residents were not to be invited to customize their spaces to suit themselves. In some cases, this attribute of modernist design may be put down to the narcissism of the designers. In many cases, however, conformity was an implication of the modernist view of industrial production as the ultimate state of civilization. The efficiencies of industrial design and production would drive out suboptimal or idiosyncratic architecture, replacing it with universal design. In that event, people would have to accommodate themselves to their designed surroundings, rather than the reverse. This view is admirable in its egalitarianism, that everyone should enjoy the same, amenable standard of living. However, it is also objectionable in the sense that it aims to achieve this end through imposition of a rigid mode of life.

In any event, models were important to the development of modernist design. The influence of ships, in particular, on Le Corbusier was noted above. The ability of passenger ships like the Aquitania to efficiently accommodate hundreds of people on long voyages clearly impressed Le Corbusier. In his architecture, he sought to apply the lessons of ship design, as he saw them, to the design of houses. The efficiency with which modernist designs could provide accommodation for large number of people, using modern, mass-produced materials such as concrete, steel, and glass, made it a highly suitable building regime for the post World War II construction boom.

Of course, modernist models had their limits. After all, ships (and cars and airplanes, other sources much admired by Le Corbusier and contemporary modernists) are not themselves machines for living but machines for transportation. The arrangements made in such vehicles for people to travel in them temporarily are not necessarily appropriate for structures where people expect to live permanently. The conformity required in arrangements for air travel, for example, is more of an imposition in a house or a neighborhood.

However, such limitations might be overlooked because industrial models also served their ideological function. That is, living in industrial surroundings would accommodate people to the industrialized world that they inhabited. In his way, then, Le Corbusier was just as concerned as Pugin for the authenticity of his architecture. However, cars and ocean liners served him as guarantors of authenticity and the good life in place of the medieval cathedrals favored by his predecessor.

4 Industrial Design: Looking to the Future

Besides looking to the past for inspiration or to the present, it is also conceivable to look to the future for models suitable for the purposes of good design. Although such a perspective may sound paradoxical or even impossible, it was characteristic of a third design movement that I wish to examine, that being industrial design.

Strictly speaking, industrial design is not a movement but a profession. It arose as a result of the industrial revolution and the mass production of goods. Before the revolution, household goods were typically produced by craftsmen who worked in local design traditions, producing a given item from raw materials. With industrialization, craftsmen were replaced by semi-skilled laborers who did not participate in the design process. Goods were designed either in mere imitation of previous crafts traditions, or they were designed by their inventors. By the early twentieth Century, both these approaches had proven inadequate for the novel technologies that were being mass-produced. A group of professionals arose whose occupation was giving proper form to these new technologies. These were the industrial designers.

In spite of the fact that industrial design was, and continues to be, a profession, its first practitioners shared a set of values that informed their work. Thus, the profession also constituted a design movement with a characteristic ideology. Perhaps the key values of this ideology were *progress* and *consumerism*.

The value of progress in industrial design was most clearly captured by Raymond Loewy, perhaps the most famous industrial designer of his era. Loewy (1893–1986) was born in France and received an education in technology in a preparatory school in Paris. He served in the French Army Corps of Engineers during the First World War. After the war, he emigrated to the United States where he made a living applying his artistic talents as a window dresser for New York department stores and as an illustrator for fashion magazines. His first break in industrial design came with the commission to redesign the Gestetner duplicating machine in 1929. Afterwards, Loewy established a successful design consultancy and participated in the design of a variety of industrial objects, from cigarette packages to cars, locomotives, and refrigerators.

In his autobiography, Loewy tries to capture some of the lessons he had learned in the course of his career [10]. One of the key lessons is embodied in what he calls the MAYA principle. “MAYA” is an acronym for the phrase “Most Advanced Yet Acceptable.” In his view, a well-designed gadget should appear to its users to be technologically advanced but also comfortably familiar. Loewy had observed a tension in people’s minds about what they expect from the things they use: On the one hand, people expect technology to improve over time, so that a newer gadget should outperform older ones. As a result, they expect the design of their gear to change over time. On the other hand, people like to stick with what they know or are used to. Thus, change in design can be discomforting or unwelcome. The MAYA principle suggests that industrial design has to balance people’s expectation of innovation with their need for stability.

Consider Loewy’s redesign of the Gestetner mimeograph machine [11]. The Gestetner was an industrial contraption with an exposed mechanism and perched on an ungainly metal frame. Loewy enclosed the mechanism in a streamlined case and streamlined the machine’s appearance and footprint. By enclosing the machine’s workings, Loewy made it less dangerous, e.g., the user’s tie and fingers were not likely to get caught in its gears and its toner was less likely to get on the user’s skin and clothing. The new appearance also made the machine more approachable.

Loewy’s redesign of the Gestetner provides a good illustration of the MAYA principle. The new design was advanced in the sense that it brought the productivity of an industrial device into the office space. Beforehand, the ungainly and mechanical look of the Gestetner had caused users to categorize it as industrial equipment. Thus, it was treated like a furnace or a boiler and hidden away from the office spaces where its duplicating function was most useful. Afterward, by making the Gestetner look and feel much like a file cabinet, Loewy caused office managers to think of it as a piece of office furniture, to be kept in the work place itself. Thus, the new design was advanced in the sense that it brought industrial productivity to the office, and acceptable in the sense that it looked right at home next to the file cabinets and desks already situated there. As a result, the new Gestetner sold well as piece of office furniture.

The MAYA principle illustrates the importance of progress to industrial design of this era. Advancement, on this principle, is an indispensible part of the design of new goods. That is, one of the jobs of a good designer is to provide customers with goods that will outperform previous designs, thus making the work of customers more productive, and their lives more pleasant. As a practical matter, the MAYA principle also instructs us that progress is best served up in moderate portions. This view stands somewhat in contrast with the view among some current designers that advancement should come in the form of game-changing or disruptive designs.

Consumerism is also part of this picture, although not one that is explicitly noted in the MAYA principle. If there is to be advancement in design, then existing designs must become obsolete. New designs can make old ones obsolete in at least two ways. First, new designs may perform a given job better than old designs. A new engine, for example, may burn fuel more efficiently than older designs. Second, new designs may appeal to people more than old one designs. The practice of changing the style of cars each year provides a good example: People may get rid of an old car in favor of a new one not because the new one is technically superior but because the appearance of the new car makes them feel unhappy about the appearance of the old one. This mental phenomenon is known as *psychological obsolescence* [12] .

One example of how industrial design could be applied to psychological obsolescence is provided by Loewy. One of Loewy's best-known designs was a streamlined pencil sharpener. The sharpener was designed in the shape of an aerodynamic tear drop, with the hole for insertion of the pencil tip at the round end and the handle to turn the mechanism at the pointed end. William Gibson describes the sharpener as follows [13]:

> The Thirties had seen the first generation of American industrial designers; until the Thirties, all pencil sharpeners had looked like pencil sharpeners; your basic Victorian mechanism, perhaps with a curlicue of decorative trim. After the advent of the designers, some pencil sharpeners looked as though they'd been put together in wind tunnels. For the most part, the change was only skin-deep; under the streamlined chrome shell, you'd find the same Victorian mechanism. Which made a certain kind of sense, because the most successful American designers had been recruited from the ranks of Broadway theater designers. It was all a stage set, a series of elaborate props for playing at living in the future.

The point is that the mechanism of the sharpener has not changed. Loewy has simply made the casing more up-to-date.

This application of industrial design encourages consumerism in the sense that it invites users to confuse technological innovation with stylistic innovation. In the case of the pencil sharpener, this confusion could lead users to dispose of their existing goods in order to purchase new ones that do not sharpen pencils any better.

Gibson also observes that industrial design of the era allowed people to “play at” living in the future. This point is key to see how the use of models fits into this version of industrial design. Designers like Loewy could not, of course, actually see into the future and take from there the models they needed for the present. They could, however, take current trends in technology and extrapolate them. One trend they could extrapolate involved streamlining, that is, the use of aerodynamic shapes. Industrial designers of that era felt that air travel was the transportation of the future for all. Each middle class family would have its own autogyro parked in its driveway, ready to fly them to work, on shopping trips, to baseball games, and so on. (In the movie *Things to come* [14], the heroes fly an autogyro in the year 2036. The autogryro in the film was designed by the industrial designer Norman Bell Geddes.) Thus, industrial designers took existing aircraft as models, imagined how they would look in future, and then applied these ideas to the design of contemporary goods, even pencil sharpeners. Thus it was that industrial designers could look to the future, as it were, for models to apply to contemporary design problems.

As with revivalists or modernists, industrial designers of the early twentieth Century used models in order to address design problems. Their models were selected not merely for their ability to answer questions of utility but because they embodied the ideals of the movement. Central to the ideology of that movement were the ideals of progress and consumerism.

5 Conclusions

Models in design can serve at least two functions. First, models may be good sources of information about how the demands of utility can be met. In this sense, models serve as “model solutions.” Second, models may be good sources of validation about how life is properly lived. In this sense, models serve as “model citizens.”

Discussions of model-based reasoning in design typically focus on models in the first sense. This focus is understandable as the value of models as model solutions is crucial to their application in design. However, models are applied by designers for reasons besides their relevance to utility. That is, designers choose models that are model citizens, that reflect and reinforce values central to their ideology. Pugin chose medieval structures as models for his architecture because they embodied his values and because the results would promote those values to his contemporaries. Le Corbusier looked to cars and ships for similar reasons, as did Raymond Loewy with futuristic aircraft.

The dimension that most clearly differentiates the three design movements discussed above concerns the attitude to history embedded in each one. Pugin and the other Gothic Revivalists thought that history had reached its apogee in the Middle Ages, with a decline, at least in Britain, after that. Le Corbusier held that history had reached its high point in the industrialized world of his own day.

Loewy and other industrial designers saw advancement as extending indefinitely into the future. Each designer naturally looked to the greatest historical epoch for models to help him address his own design problems.

Most likely, the same can be said of any professional designer. As they seek out models to assist with design issues, they look not only for instruction but also for validation.

References

1. Tweney, D: iPhone's design tribute to a 1977 Braun calculator. Accessed 2011; 5-Dec. from Wired: http://www.wired.com/gadgetlab/2007/07/iphones-design/ (20;July)
2. Isaacson, W.: Steve Jobs. Simon and Schuster, New York (2011)
3. Aldrich, M.: Gothic sensibility: the early years of the Gothic Revival. In: Atterbury P. (ed.) A. W. N. Pugin: Master of Gothic Revival. (pp. 13–30). Yale Univeristy Press, New Haven (1995)
4. Saint, A.: Pugin's architecture in context. In: Atterbury, P. (ed.) A. W. N. Pugin: Master of Gothic Revival, pp. 79–102. Yale University Press, New Haven (1995)
5. Hill, R.: Augustus Welby Northmore Pugin: a biographical sketch. In: Atterbury P. (ed.), A. W. N. Pugin: Master of Gothic Revival, pp. 31–44. Yale University Press, New Haven (1995)
6. Meara, D.: The Catholic context. In: Atterbury, P. (ed.) A. W. N. Pugin: Master of Gothic Revival, pp. 45–62. Yale University Press, New Haven (1995)
7. Corbusier, L.: Vers une architecture. Vincent, Fréal & Cie, Paris (1923)
8. Curtis, W.J.: Le Corbusier: ideas and forms. Phaidon, Oxford (1986)
9. Rybczynski, W: High hopes. In: Rybczynski, W. (ed.) City life: urban expectations in a new world, pp. 155–172. Scribner, New York (1995)
10. Loewy, R.: Never leave well enough alone. Simon and Shuster, New York (1951)
11. Barmak, S: A pioneer of user-friendly. Accessed 2011; 5-Dec. from Toronto Star: http://www.thestar.com/sciencetech/Ideas/article/238172 (2007; 15-July)
12. Slade, G.: Make to break: technology and obsolesence in America. Harvard University Press, Cambridge (2006)
13. Gibson, W.: The Gernsback continuum. In: Carr T. (ed.) Universe 11 Garden City, pp. 81–90. Doubleday & Company, NY (1981)
14. Wells, H. G. (Writer), & Menzies, W. C. (Director): Things to come [Motion Picture]. United Artists (1936)

How Affordance Changes: Based on the Coincidence Model of Technological Innovation

Cao Dongming, Luo Lingling and Wang Jian

Abstract Affordance is a concept that was first coined by perceptual psychologist J. J. Gibson over 30 years ago and now has been widely used in designing man–machine products and systems. The central question we ask is "how does affordance change during technological innovation?" Thus, this paper will examine real technological innovations in order to pursue the trail of affordance, finding out where it comes from and how it changes throughout the process of technological innovation. Technological innovation, which is a creative and dynamic process, forms a chain, and each link of this chain has an interaction between different products. In this respect, the chain acts as an *affordance chain.* The affordance chain not only provides an interface between designers and machines, but also an interface between machines and users (or consumers). The success of innovation depends on whether consumers (or users) accept the design or not and whether a specific set of criteria was followed. Therefore, affordance helps us to understand and explain the link between designers, products and users. Moreover, as technological innovations experience both the process of transverse development and historical evolution, a two-dimensional space is needed to describe their position. For that reason, affordance should also exhibit some kind of evolutive accumulation.

1 Introduction

The term *affordance* was first introduced by psychologist James J. Gibson in 1977 through his paper "The Theory of Affordances" [1]. Shortly thereafter, in 1979, Gibson further developed the theory in his book *The Ecological Approach to Visual Perception* [2], which established him as one of the most important

C. Dongming (✉) · L. Lingling · W. Jian
Department of Philosophy, Northeastern University, P.O. Box 229 110004 Shenyang, China
e-mail: cdm0429@163.com

L. Magnani (ed.), *Model-Based Reasoning in Science and Technology*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 8,
DOI: 10.1007/978-3-642-37428-9_34, © Springer-Verlag Berlin Heidelberg 2014

cognitive psychologists in history. Gibson's theory not only explored a new area of interest for such fields as cognitive and environmental psychology, but also assisted those working in industrial design, HCI, interior design, and Learning Study. It gradually became a basic concept for understanding the relationship between humans and their environment. Still to this day, the implementation of affordance can be seen in the literature of many fields, particularly the philosophy of technology.

However, there appears to be a gap between the term's original meaning as defined by Gibson and its later interdisciplinary application. First, Gibson's theory of affordance mainly discusses how animals perceive their surroundings. Thus, experiments for this theory were primarily performed with the environment in its most natural state. Only a couple of relatively simple items were included in the experiment, such as a ladder and a chair. We can then argue that an explanation is needed to clarify how the application of affordance may transition from a simple, archaic, and individualized natural environment to a complicated, modern, and collective humanized environment, especially, as the philosopher Jacques Ellul branded, to this technological-society environment. Secondly, Gibson's theory largely focuses on the relationship between animals and the overall environment encompassing them. While later studies, like design study, chiefly explain an individual (person)-to-individual (product) relationship. With this type of application, we can also argue that an explanation is needed to understand the difference between integrity and the particularity of information acquired from affordance. Third, Gibson's theory is initially placed on an evolutionary scale, while later, it is used to explain a moment-to-moment relationship.

Therefore, we are faced with the challenge of linking original meanings with later applications in order to address such pressing issues. In order to accomplish this task, we need to answer the following questions: In today's real and humanized world, has affordance ever changed? If so, how has it changed? What caused it to change? What has resulted from its change? To answer these questions, we must continue to explore this theory and examine its application within multiple fields.

In this study, we have avoided semantic and philosophical debates so as to closely observe how affordance performs during specific activities of technological innovation. Through dissecting the technological innovation process, we can analyze how affordance exists and develops. We set "how affordance changes during the process of technological innovation" as our primary question for the reasons stated below.

According to Gibson, "[t]he affordances of the environment are what it offers the animal" [3]. Likewise, as emphasized by Costall [4], it is only logical that the technological environment or the artificial environment (also known as Humanized Nature [5]) should have affordances as well. Gibson continues, "An affordance is neither an objective property nor a subjective property, or it is both if you like. An affordance cuts across the dichotomy of subjective–objective and helps us understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither; An affordance points both ways, to the environment and to the observer" [6]. If such reasoning is to be believed, then affordance can be viewed as a new way to study the development and the evolution of technology because it is the technology that can be serving as a bearer of such relationship.

Second, since affordance is able to inspire and produce corresponding behavior in animals, we are able to observe how affordances arise, develop, and evolve from both technological creation and innovation. It is with human technology that a particular behavior will be displayed or a product will be accumulated. Inevitably, there should be traces of affordance. In this view, if the process of developing and innovating technology can be clearly described, then it is possible more affordances will appear.

Third, the process of technological innovation was selected as the object of analysis because the process of technological innovation is more complete than the process of technological invention and technical design. Stroffregen adds that the inherent economic connotation within technological innovation, which makes affordance and it's "mechanism of initiating an action" [7] may exhibit a relationship with such characteristics as ethics, values, cultural, practices, and even religion. In addition, what is the true relationship that exists between the invariant features of affordance and its evolution? If affordance remains constant, the relationship between people and the environment may become apparent, and the process towards complete technological innovation, causing "human cognition and a transformed world" may begin.

Therefore, the most important question we must ask is, "how does affordance change during technological innovations?" For this reason, we need to revisit real technological innovations to track affordance, finding out where it comes from, where it goes, and how it changes during the dynamic process of technological innovation.

2 The Coincidence Model of Technological Innovation

Technological innovation is an evolutive process with multi-dimensions. This paper tries to clarify the complicated picture of this process. It involves two-dimensional technological processes and three-dimensional demand processes. In addition, it asks how such technological aims can meet these demands. By completing a thorough historical and macro analysis of their relationship, we can offer suggestions as to how innovation may become more efficient. We will focus on coincidences that occur in time and space for technological innovation and the demands of such relevant innovation.

In short, if ever there were a model for technological innovation, it should reflect all the characteristics we have mentioned.

First, as the success of any technological innovation depends on a direct or coincidental "encounter" of technological opportunities and market demands, the model embodies the *purposiveness*, *causality*, and *contingency* which are contained in the process. (1) Purposiveness is the motor that runs the process of innovation. It is the motivation and goal of technological innovation. The existence of intentional responses to stimuli or tropism, determines the validity and value of causality. The uncertainty of the destination in the time domain is the core reason for contingency. It is this intentional tropism that unites causality and contingency in functioning towards technological innovation. (2) Causality, on the other hand, serves as the base and prerequisite of technological innovation. If causality did not exist, then there would be no understanding of technological activities and no knowledge on how to organize them. (3) Contingency is the necessity of technological innovation. It is with contingency that technological innovation is equipped to be uncertain, scarce, and even a bit secretive. Therefore, the success of technological innovation will always be attractive, desirable, and lucrative.

Second, the model presents under a broad perspective. It breaks the line separating internal and external technology, and reflects the integration of the technological innovation process. The model also reveals the dynamic and evolutive aspects. It tries to locate a concrete innovation within a coordinate system that not only contains transverse dimensions with which the innovation actually extends along, but also historical dimensions that have been articulately developed by the technological innovation and its corresponding industry. The model should point out the social construction of technological innovation, including both objective constructions and subjective constructions.

Third, the model tries to be symmetrical and impartial in its style of explanation. In other words, if we helped promote technological innovations that we believed were good for the world, then simultaneously, we should try to prevent or regulate technological innovations that we believe would harm the world.

In the analysis above, we attempted to present a concrete model that depicts the abstract concept of technological-innovation success (see Fig. 1). The model shows the materialization of technological opportunities that result in the innovation of end products. It also shows market opportunities resulting from current consumer demand, which then interacts with various forms of the dynamic process. Below, we also provided a detailed explanation of the model.

Within the model, there exists a light gray area that represents the so-called project field. This region stands for a specific, technology-related industry within time and space. It is dependent on technological innovation and the existence of social demand for a specific technological innovation, reflecting the trend of the need for the technological-innovation process.

Need, however, exists on both sides, which then can be divided between a technology-based group and a social demand group. Within each group, there exist individual features that affect each, such as the location of such moving tendencies (excluding the center) wishing to connect and the direction within the field. At the same time, features showing the process of interaction on each level must not be ignored. When many features are gathered together, a core group is formed. The small core group (dark grey) and its peripheral group (gradually fading color) have similar structures. The similarity embodies the process, permitting others to

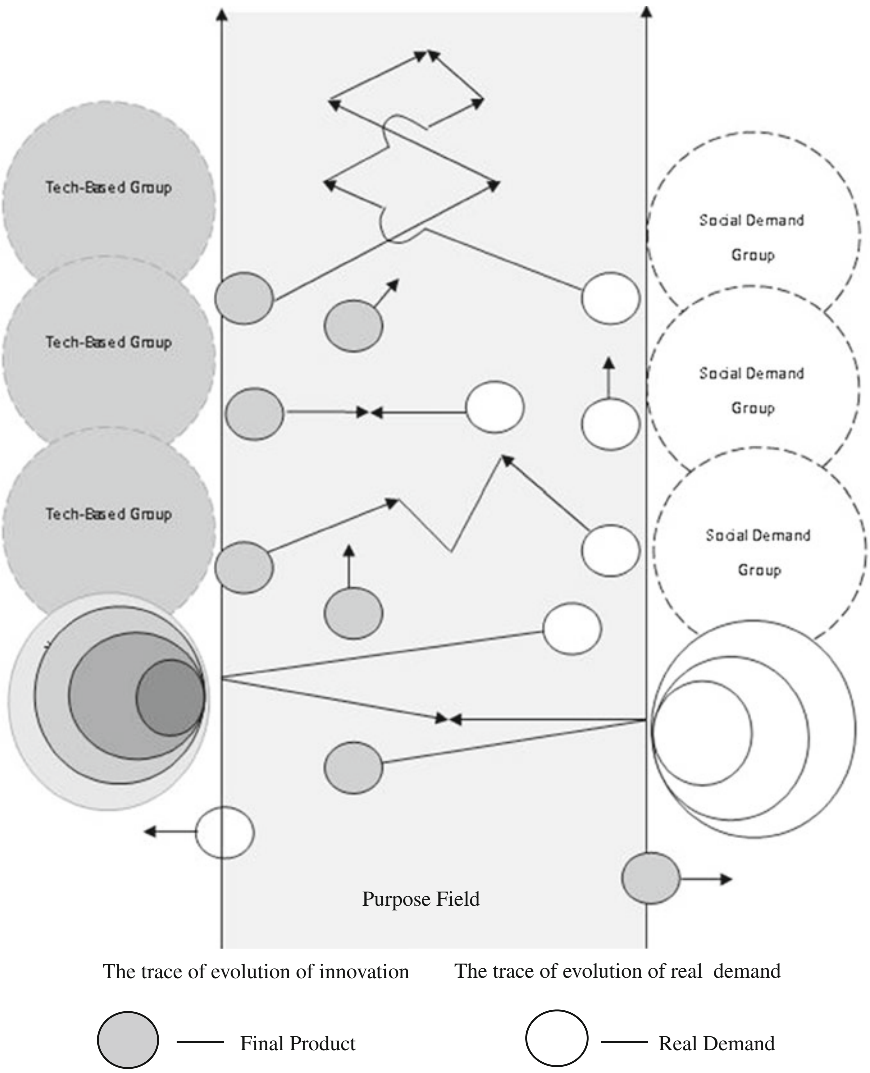


Fig. 1 Coincidence model of technological innovation

acquire certain aspects or "shades" of it. It is the quintessential embodiment of the technology innovation process.

The purpose of the technology group lies in its tendency for traction, which is derived from the law of causality for performance technology to complete the cycle of innovation. If the industry is chemical manufacturing, for example, then decisive quantitative research needs to be conducted first, which is then followed by production. The end product shows the essence of original product, which was presented during the initial stages of when the idea for the technological innovation was first formed.

Let's begin by examining the outer layers of the demand group. The first level represents universal human needs. It stands for the basic necessities required to physically sustain human life, such as food, clothing, housing, and safety. It also includes social needs, such as loyalty and righteousness. In addition, it incorporates personal needs, such as knowledge and self-expression. When all three aspects of human needs (physical, social, and metaphysical) are combined together, they form the basic component of human nature. The second level is a representation of desire. It identifies a human's drive to shape his or her own culture and character. For example, in the United States, a person may consider a hamburger, French fries, and a bottle of coke as a normal meal, but in Bali, another person may deem pork and peas as the go-to food. Desire that satisfies one's demand can also physically describe that same individual. The desire of a human is nearly infinite, but the resource to supply such desire is finite. Thus, people must use a finite amount of money to choose those products with the greatest value and satisfaction. When a masterpiece is backed by a purchase, then desire becomes a demand. This type of demand belongs to the group's third level. Consumers are met with all types of products catering to their interests; however, they will only choose those products with the best prices and most compatible features. For instance, we can examine Honda, a Japanese automobile manufacturer. The Honda Civic belongs to their line of subcompact and compact automobiles, which is a popular mode of transportation. Paired with having a low price, the automobile is very fuel efficient and easy to maintain. Since the date it was established, Honda has continued to provide, to the best of their abilities, satisfaction to those with an interest in their products.

Examining the longitudinal axis, we can see that the process of technological innovation has been changing and developing throughout space and time. We can also see each technology-based group and the degree of demand for that group over time. Finally, we see the group's evolution and development. As previously mentioned, the platform for technological innovation and the formation of market demand levels are constantly evolving and firmly remain a dynamic part of the process.

The gray and white ball, as already stated, represent the end product of innovation and the current consumer demand, respectively. They move within the project field. Different shades of each group may appear, and different paths of collision frequently occur in order to successfully meet a new technological innovation. Sometimes a ball may move parallel or cause a technological innovation to fail due to its trajectory being indefinite. On the whole, technological innovation is based on random chance.

In the chart, each gray ball contains specific levels and depth. The outer layer is the most shallow (represented by a lighter shade of gray). It shows the vulnerability of the system and how it lingers on the edge of uncertainty. Only time will reveal exactly what elements may enter into the process of technological innovation as these elements are correlated. The so-called technological ties that people often speak of actually have nothing to do with the elements. After all influential elements have been integrated, they are then represented by an object-path line, indicating the development direction of an industry.

3 The Modularization of Affordance: With the Developing Process of Technological Innovation

As a common view, technological innovation is the realization of an idea for an innovation, which we call a dynamic process. During this process, the idea will be expressed as in several object forms, like a blueprint or model, sample product, mass production product, or commodity. The different subjects within this process, like designers, engineers, and managers, are able to co-operate with each other by "picking up" any affordance that appears from objects of different forms. There are usually a group of innovations that join together to achieve an original idea. This is a process of accumulation, construction, and modularization of different affordances.

3.1 The Accumulation of Affordance

We will take the Model T automobile, which was designed and developed by Ford in 1908, as our example. Ford originally designed this automobile under the belief that an automobile "should be more in quantity, better in quality and cheaper in price than any other automobile". As repeatedly asserted by many biographical authors of Ford, we can understand that this idea was a reflection of Ford's distinct personality [8–10]. In order to realize his idea, Ford made strategic technological decisions based on the overall technology level of North America during that time. He applied standard measures of production, and he only manufactured single and economical models. Next, he set greater defined technical indexes. The fuel consumption rate was set at 20 miles per gallon. The engine was to follow a simple design. This automobile was to be of a higher class than other vehicles, and other features of this nature were defined. Producing a "high class" was a primary goal as, during this time, many rural areas had rough, unleveled roads. Outside of the city, the conventional way of creating a road was to spread soil two inches deep, particularly along a narrow path or farmland. Therefore, roads were extremely dusty. The Model T automobile was designed to be high from the ground with a durable structure and net space. Its ability to adapt to roads of any conditions gave it a compelling advantage over other models. Even bumpy roads around field grounds were no problem for the Model T. The Model T raised the bar for the automobile industry, forever changing the way people viewed them. We can also look to the United States for more examples. In 1890, it produced one-third of the world's steel, directly causing the rubber industry to hit its prime. In the 1850's, tires were used for carriages. Although the theory of an air valve existed a half a century before August Schrader applied it to rubber products, his innovation revolutionized the market. The United States paved the way, becoming the leaders of manufacturing and technology. Mass production became key for the production of machinery. Personnel were trained in mechanical engineering, manufacturing, and technology so that they could apply their technical background to further advance the industry.

After carefully considering this example, what inferences can we extract? In fact, all of these are technological expressions of affordance that the environment of the time offered to the innovators. In other words, the technological goal is actually the quantified idea of the innovator's idea, resulting from the combination of purpose, technological level of contemporaries, and future technology to be used under objective environmental factors. Its essence is a linguistic expression that Ford and his engineers "picked up" at that time, providing the environment of the automobile and its usage with the affordance of such technology. The affordance in North America at that time is considered to be the technological environmental decision that was perceived by Ford. The language provides a unique way to express the translation of a specific and technical language form. Briefly speaking, all these factors are the resulting affordance made from the decisions of innovators. After combining such factors, technology was formed, producing a more delicate and precise technical language expression.

3.2 The Modularization of Affordance

The structure formation of the affordance process modules defines the modularization of affordance. Affordance is gathered in order to process, combine, and screen. Further screening is completed to determine its "structure of choice". In the case of Ford's Model T, the process of technological innovation is the combination of such different subsystem sectors as production innovation, production technology innovation, and market innovation for affordance. The three divisions of the structure constitute the core layer of technological innovation. Product innovation involves creating a new model, such as the Model T. Production technology innovation requires setting up an assembly line or something similar, following the auto industry as an example. Market innovation indicates initial signs of sales and the direction of its network. Ford's network reached all over, touching all corners of the earth. Hard work was needed to inform as many people as possible about the Model T as well as providing them with a way to directly contact distributors of the product. Dealing with consumers directly was the best approach to cultivate demand for this technological product. The logic behind these three divisions of innovation can be summarized as follows: First, product innovation provides a clear and concise point of entry for prominent automobile products. Second, market innovation increases awareness about the Model T, informing consumers of its competitive advantages as well as promoting consumer group expansion. Finally, production technology innovation efficiently helps expand any given market to meet personal and creative product needs as well as preserving the integrity of the achievements of the innovation. The affordance module is equipped to handle everything, from the Model T to all innovations. Once the module is formed, future similar innovations, often called *modular packs*, can be applied, becoming "new" in terms of affordance. Network marketing, which once is born, will continue to spread following the trend of time. The reason for this new affordance may lie in its efficiency. Ford's sales network covered vast areas. The huge amount of orders—real demands—further refined production methods of innovation as well as "chance" and the "opportunity to a trigger a new mode of production". As a result, industrial production was born, standing as a symbol of production mode, or more specifically, the assembly line production mode. All three technological innovations are mutually supported to form one type of internal organic structure. Finally, the entire technological innovation is combined with the affordance *module* to form a more complex module. Once the affordance module is functioning, then similar practices (conventions) take effect.

3.3 The Stabilization of the Affordance Module

In order to stabilize the affordance module, a new *niche* is formed. Ecologists long ago introduced the concept of niche, which is the habitation of an organism at a specified interval. Animals are thought to use or occupy a specific environmental niche; however, this is not the same as a population's *habitat*, which refers to how (instead of where) a population lives. Gibson believes that "a niche is a set of affordance" [11]. For instance: in an advertisement for Lexus (which is a luxury automobile produced by Japanese automaker Toyota), the automobile's power, performance, and speed are tested. In order to present its smooth and steady riding ability, the hood of the luxury automobile is stacked with glass, while being placed into a high-speed simulator. The audience witnesses the glass remaining silent and still throughout the simulation. After combining the affordance module with other modules, an innovatively new and high-quality automobile is produced, offering a great experience for users. In the final stage of technological innovation, in order to inform outsiders of the maker of such new technology and increase its acceptance, which are the major goals of this stage, the new affordance module must be commensurate with the efforts when providing a narrative. This may explain why, from creation to production and from production to release, it is so costly in time and money. In the United States, in 1990, advertising firms spent over $1.74 billion on advertisements in newspapers and magazines, $230 million for radio advertisements, and more than $36.85 billion on television advertisements [12]. If information is accurately conveyed through advertisements and certain unknown technology is able to resonate with consumers, then they will be convinced of its potential and willingly consume. The moment innovative products gain recognition by many consumers, which is also when a new ecological form is brought into existence, then the technological innovation will have been met with success.

In summary, the objectives of technological innovation will consist of a technical analysis of the initial idea as to become part of the technological innovation group. Afterwards, various types of affordance will be revealed, gathering around this objective. If the time is just right, then a technological feature system will emerge. At this time, the affordance for the technological innovation will circle around the relevant affordance group. From the outside looking inward, we will see many related elements gathered around the objective, which is the result of this open and self-organizing system. The technological objective of production will be realized when there is an internal demand for the technological innovation. However, unlike simple inventions, commercial technological inventions, as well as technological innovations, require a specific, economical demand. Conversely, the only way to pursue economical and technological objectives is to create a technological innovation. Furthermore, this process devises a way to mass-produce a single technological innovation. Accordingly, the objective is again revised. Again, if the timing is right, an industrial system will appear. An industrial system, which is also a self-organizing system, will begin producing. Once a batch of innovative products is created, it will face several tests. The first test compares the product to the original idea and decides whether there is sufficient affordance to prevent it from disposal. The second test questions whether the product is still a complete representation of the original ideal and whether it can resonate with the user. In other words, it questions whether consumers are able to "see" the product in their minds and whether the product's affordance has been "picked up" (Gibson used this verb to express the perceiving action), thus, being able and willing to consume. This will also determine the final success or failure of the innovation.

Therefore, from the initial idea of the innovation to quantifying the physical and chemical process, affordance is gathered, organized, and presented in a module to process and form new affordance. This new affordance, with all of its integrated elements, represents the original affordance in its natural state. These "new" features depend on their close connection to affordance that has been combined with its objective, thus becoming a tightly packaged convention or common practice.

4 A Niche's Evolution: Trailing the Evolution of Technological Innovation and Affordance

Regarding technological development, the Chinese philosopher Chen Changshu describes this area as having four distinct characteristics: the technological development of self-growth, the technological construction of compatibility, the gradual technological transitions, and the measures of rise and fall of the technological evolution cycle [13]. James M. Utterback comments that the industry development is basically composed of the “fluid phase, transitional phrase, and specific phase” [14]. However, in this article, we will not focus on any certain phase. Technological development is represented throughout history. For example, it can clearly be seen in the development of the automobile industry. Strictly speaking, technological innovation and industry are closely related. Sometimes, a key technological innovation will trigger the birth of an industry. Likewise, as industry develops and evolves, technological innovation will reveal different features. As the affordance of technological innovation changes during the evolution process, evolution features are presented. Evolution provides a way for affordance to gather and for a niche to appear, grow, and stabilize. The mechanism of evolution lies in the bootstrapping method of accumulating affordance, particularly by passing on such inherently redundant features.

4.1 The Appearance, Growth, and Stabilization of a New Niche

First, a new niche is formed. As we previously said, Ford’s Model T automobile held the leading position to revolutionize the automobile industry in the United States, which before that time, had remained in the fluid phase. It was also during that time that product innovation took the lead. The fluid phase involves many moving components, causing the results to be ambiguous in areas such as production, processing, structuring, and management as well as obtaining the leading position among competitors in any given company. Currently, over a hundred enterprises are involved with automobile manufacturing as product innovation is facing uncertainty in terms of its targets and technology. Due to the rapid evolution of products, technology is hastily developed for new products, which are often sloppy, expensive, and unreliable. The preliminary stage of the fluid phase for processing innovation is normally incomparable to that for product innovation. Second, a new niche is grown. The era of Ford’s Model T brought in the transformation phase of the automobile industry for the United States. When the market for a new product grows, the relevant industry enters into the transformation phase, as defined, which is reflected in the recognition of product innovation by such market as well as the leading design. As more knowledge is gathered about a user’s needs, customization of products becomes key in competition, when companies shift their focus from innovator workspace to production workspace and mass-produce new products. During this phase, the correlation between product and processing innovations is very important. Specialized materials are used, such as installing special equipment in plants for partial automation. Administration and control suddenly play a more prominent role. As demonstrated by the ongoing rigid growth of production operation, consideration that is taken into the product evolution requires a higher cost. Third, a new niche is stabilized. Later on, General

Motors adopted an innovative market segmentation strategy in 1920. They surpassed Ford and became the new leaders of the automobile industry for the United States. As a result, the industry reached the specific phase. Processing and product innovation became the leading force. Products at the specific phase are explicitly defined, narrowing down the difference between competitors. They share a lot in common. Using the automobile as an example, we can see that they are very complicated products. Yet, they also have a tendency to adopt the same design and manufacturing solutions with the same or similar aerodynamic shape, engine, and internal design. Thus, the relationship between products and processing is very close. It is difficult to make any minor modification to products and processing, and the expenditure is certain to be considerable. Furthermore, if modification is made to products or processing, then the other needs the same modification, even if the modification is minute. For example, the restructuring of processing steps on a production line will also be considered as revolutionary modifications to the manufacturing sector. The days when inventors were in charge are gone. During this phase, monitors patrol. They supervise to make sure the production runs smoothly. This not only refers to workers but also personnel that report the work status and workers' performance. It also includes the managers and engineers that have had to change their role as a result of modifications brought about through technological change. At this time, the product innovation strategy that was adopted by General Motors seems to keep all competitors at bay. Nevertheless, the development in such industry will continue to bring forth the advent of a new phase.

4.2 A Niche's Mechanism of Evolution: Advancing Affordance's Bootstrapping Method

Gibson states that "affordance's environmental existence will only be revealed when an animal's activities within an environment are examined" [6]. From this we may understand that the majority of affordance is "below the surface". If we do not look for it and extract it, then it never would be revealed. Gibson also writes, "affordance points in two directions: towards the environment and towards the observer" [6]. As a result, the affordance supply for humans is directly correlated to the evolution of human activities and the mechanism of evolution. After a long-time, as human activities deepen in complexity, affordance will slowly and unintentionally expose itself. Affordance that is superficial may also become visible. For example, wood can be rubbed together to create friction, which produces heat and subsequently fire. This was an accident or circumstantial chance that humans learned to master during the early ages, which furthered their development. However, some layers of affordance are very deep, such as the affordance found in chemistry. This is found during the early ages when people practiced alchemy. The accumulation of affordance is found in succession. One affordance triggers another, and then another, until the whole sequence of correlated affordance is extracted. A set of affordance forms a niche, which, as we mentioned before is different than a habitat. This human activity objective is to make mutual connections in order to form a niche. This is the way in which affordance reveals its construction and itself. Thus, affordance proceeds to adopt a self-generating or bootstrapping accumulation method. Affordances are captured and used in this affordance equipment, providing further revelation.

4.3 The Features of a Niche's Evolution: The Affordance's Redundant Inheritance

As previously discussed, the evolution of an affordance's module and its niche is based on a simple combination mechanism. Namely, during every situation illustrated below, the actions were driven ahead by a basic, dynamic force, gathering a certain number of affordances and forming a combination. However, it is precisely during this type of combination mechanism that the potential to transform in the future will be exceedingly high. In simple, mathematical terms, generally speaking, for any N possible elements, we will receive 2N in possible combinations.

Suppose that a niche only contained five kinds of affordance: A, B, C, D, and E. If each affordance can only be used once, how many different types of combinations can we create? After combining A to E (or plugging in 2^5), we realize there are 32 possibilities. Similarly, for 10 kinds of affordances, we will have 1,024 (or 2^{10}) combinations. For 20, we have 1,048,576 (or 2^{20}) possible combinations. For 30, we have 1,073,741,824 possibilities. For 40, we have 1,099,511,627,776 possibilities. This type of combination probability index method varies (such as using the 2N change method). For any given number of affordance, the combination possibility is finite. When there are few in number, they will seem small. However, after becoming exceedingly small, they will immediately expand and become extremely large.

In addition, if we are able to permit the possibility of redundancy among the combinations, such as allowing ADDE and ADEE to appear as a combination in our previous example, then we will be able to see more astonishing quantities. Furthermore, in reality, redundancy is more common and more natural in general. The problem is that this type of redundancy can cause a module to change until a new niche is formed. For newcomers to this field, the redundancy process is likely to be forgotten. Thus, in order to prevent the loss of this redundancy information, we must study hard to be adapt to evolution. Finally, along with the evolution of the niche, redundancy should be piled together as to form a new environment (as compared to the original, natural environment). Chance will lead us to the path of affordance in isolation. Whether this path leads us to good or to bad, nobody knows.

5 Conclusion and Discussion

From a panoramic view, we can conclude that the change of affordance is seen in two dimensions. The first is the emergence of modularization during the development of technological innovation, which includes assembling, constructing, and modularizing the "new" affordance. The second is the niche's evolution through innovation that unfolds through its formation, its growth, and its solidification as a new niche (or a set of affordance, according to Gibson).

Gibson (1979) describes affordance as a "changing invariance" [15]. Through the perspective of technological innovation, we can see the aforementioned affordance invariance. Affordance is not dependent on the object's existence. It is its own constant. As Gibson continues to write, "The organism depends on its environment for its life, but the environment does not depend on the organism for its existence" [6]. It is said that affordance constantly changes. The "veil" is constantly lifted as to provide us with a new look at the process. Affordance is collected. Its structure is then modularized, and its niche is renovated. Gradually, the environment for human survival begins to form. The depiction of the evolution of humankind is slowly revealed.

More over, redundancies will arise with change, recursively. First, the whole process will occur on the interface between innovators and consumers as well as on every step of every phase as the entire process moves forward. In another words, there are many "mini" coincidence models that are hidden in the one we described above. Next, the redundancies of affordance will be formed as innovations are processed. After repetitive recursive modularization, a new set of affordance (or a niche) will become a redundancy.

As Gibson said, "For terrestrial animals like us, the earth and the sky are a basic structure on which all lesser structures depend. We cannot change it. We all fit into the substructures of the environment in our various ways, for we were all, in fact, formed by them. We were created by the world we live in" [6]. Thus, when discussing how to use affordance during a product's design and construction, it is important to keep an eye on the following issues: How do we deal with redundant occurrences of affordances that arise from the modularization of affordance? What is the quintessential niche? Do the natural and the non-natural roles matter to us? What niche is actually suitable for humans? Do humans need to be the driving force as to influence the change of affordance?

References

1. Gibson, J.J.: The theory of affordances. In: Shaw, R., Bransford, J. (eds.) Perceiving, Acting, and Knowing: Toward an Ecological Psychology, pp. 67–82. Lawrence Erlbaum Associates, Hillsdale, (NJ Inc. 1977)
2. Gibson, J.J.: The Ecological Approach to Visual Perception, pp. 7–129. Houghton Mifflin, Boston (1979)

3. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, p. 127 (1979)
4. Costall, A.: Socializing affordance. Theory Psychol. **5**(4), 471–474 (1995)
5. Chen, C.S.: Introduction to the Philosophy of Technology. Science Press, Beijing, p. 36–53 (2012)
6. Gibson 1979, op. cit., p. 129
7. Stroffregen, T.A. Affordances and even movement & perception. Ecol. Philos. **12**(1), 1–28 (2000)
8. Wang, W.: Qiche Shenhua (Auto Myths). Peking University Press, Beijing, p. 8 (1998)
9. Ford, H.: The Complete Works of Henry Ford. Reform Press, Beijing, p. 78–79 (1998)
10. Pollard, M.: Henry Ford and the Ford Corporation. World Publishing Corporation, Beijing, p. 67 (1997)
11. Gibson 1979, op. cit., p. 128
12. Wang, op. cit., p. 121–122
13. Chen, op. cit., p. 136–159
14. Utterback, J.M.: Mastering the Dynamics of Innovation. Tsinghua University Press, Beijing, p. 116–118 (1999)
15. Gibson 1979, op. cit., p. 7